Bayesian deep learning: Overview and challenges

Julyan Arbel

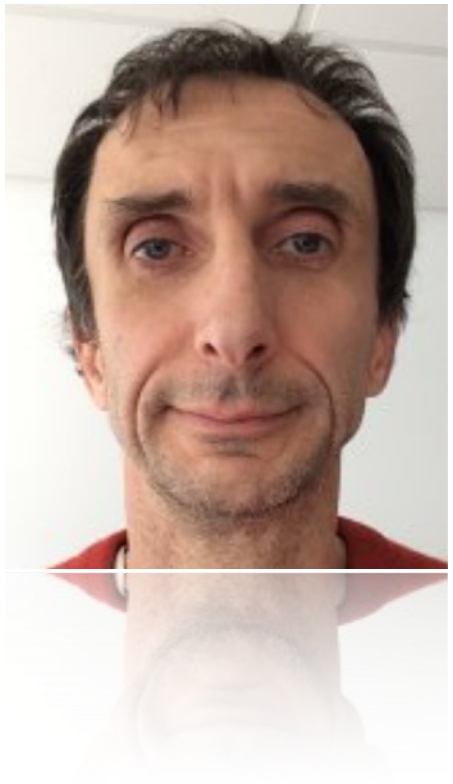
Bayesian deep learning
Overview and challenges

Julyan Arbel, Inria Grenoble

julyanarbel.com

BNP 14, UCLA, June 2025

Mariia Vladimirova
Criteo AI Lab
Paris



Stéphane Girard
Inria
Grenoble

Kostas Pitas
Cognex
Switzerland

Vincent Fortuin
Helmolz AI
Munich

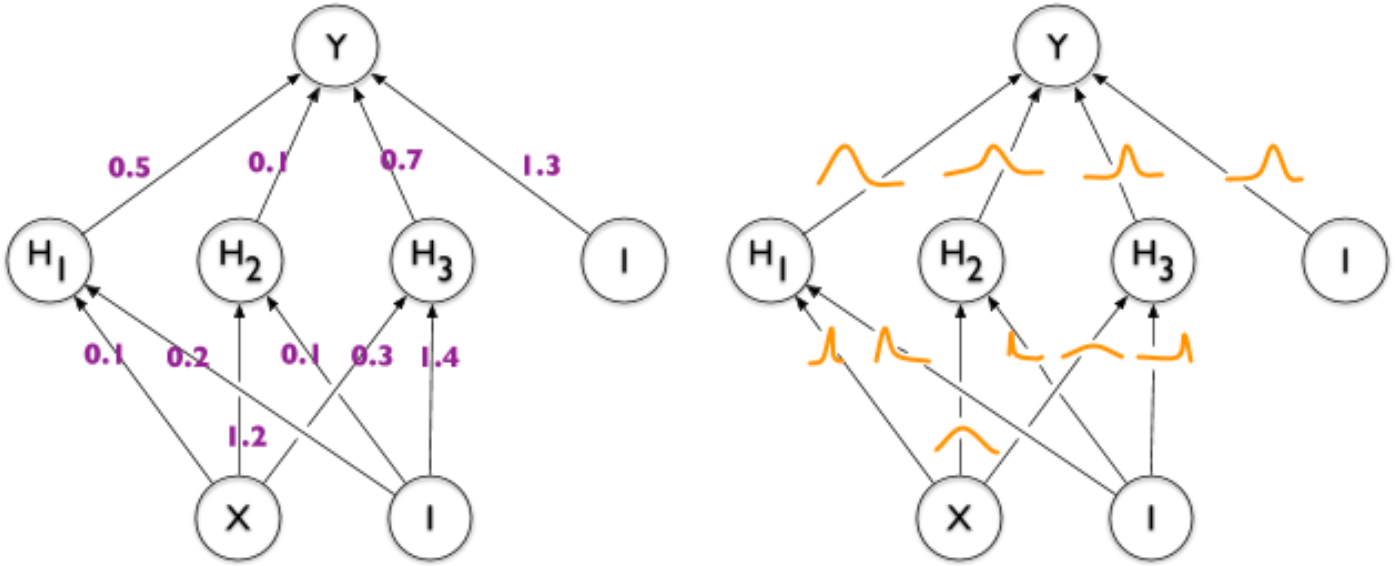
Large-scale deep learning

Considerable success ➔ 2024 Physics & Chemistry Nobel Prizes, 2018 Turing Award.

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

Bayesian deep learning

Emerged in the 90s: works by Neal and Mackay on **Bayesian neural networks**
➔ Addressing deep learning challenges by merging the advantages of the robust mathematical foundations of **Bayesian inference** with the practical effectiveness of **deep learning methods**.



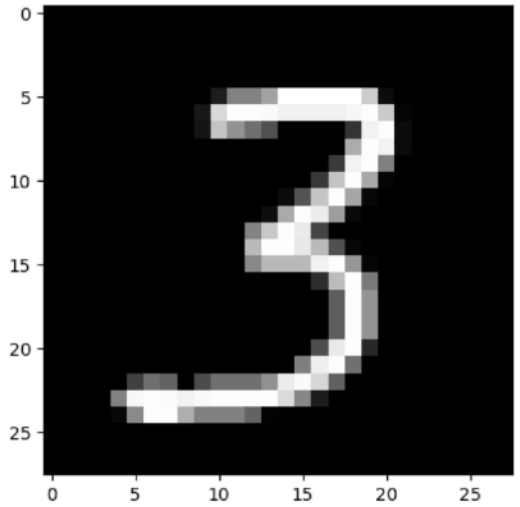
Source: [The Very Basics of Bayesian Neural Networks]

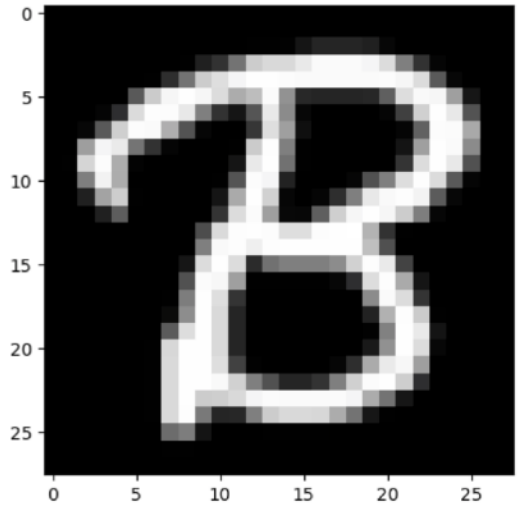
Bayesian neural networks: out-of-distribution

BNNs are simply NNs that can respond “I’m not sure”

Why is it important? Need to consider out-of-distribution (OOD) samples.

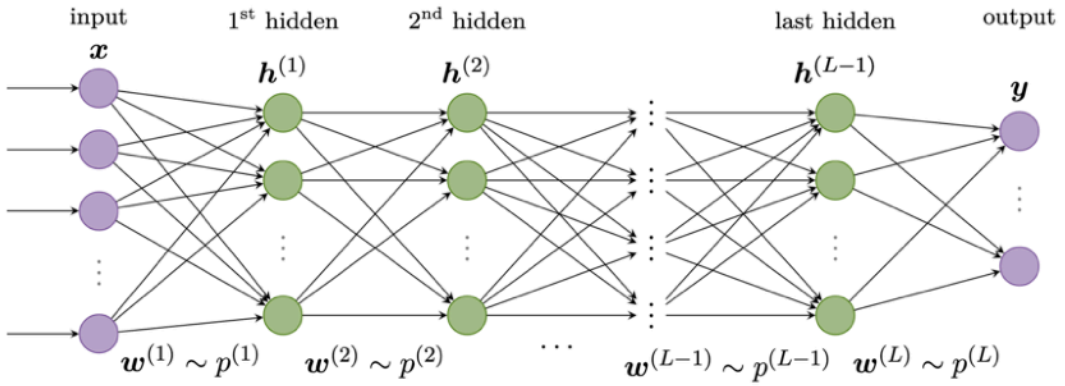
For example, train a BNN on MNIST, and ask to predict on letters instead of digits.



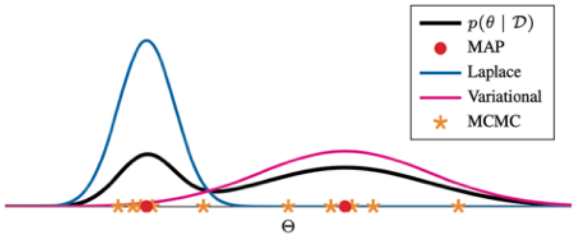


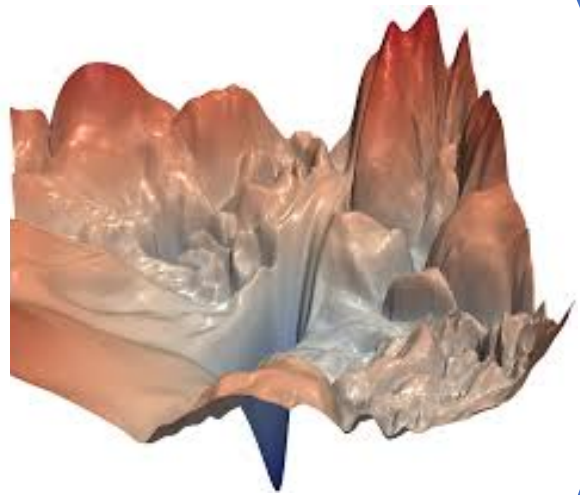
	Prediction	Probability	Std. dev. (uncertainty)
Digit “3”	3	100%	0
Letter “B”	8	57%	0,45

Priors



Posteriors





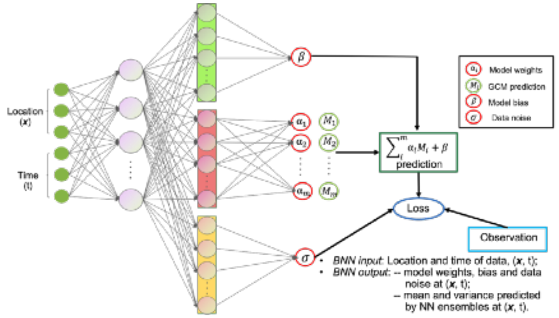
Applications

Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

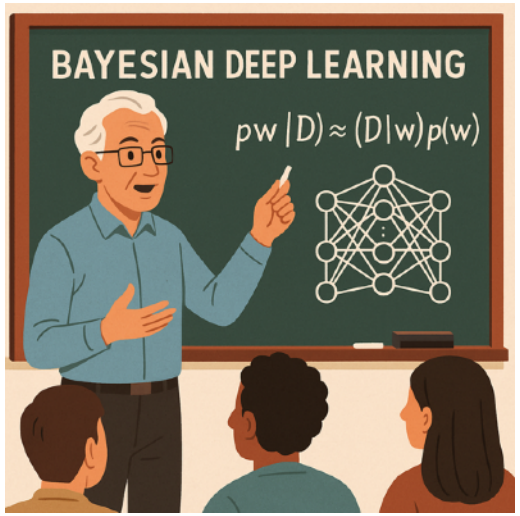
Correct answer wrong, it is "sulfuric acid"

Bing Chat Yes, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

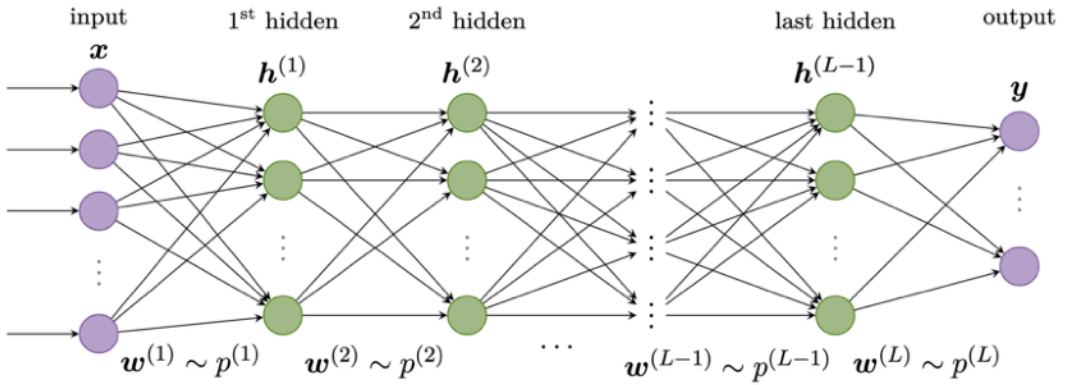
LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...



Teaching



Priors



Posteriors

Applications

Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

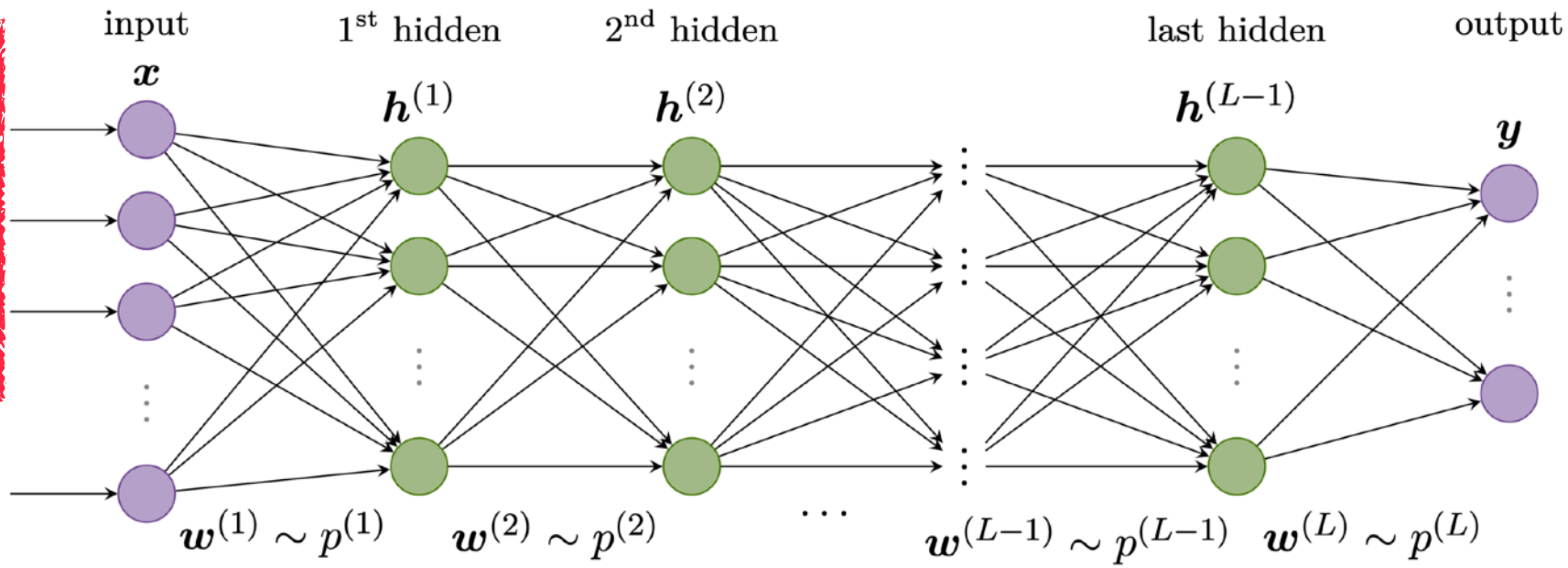
Correct answer **wrong**, it is "sulfuric acid"

Bing Chat **Yes**, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

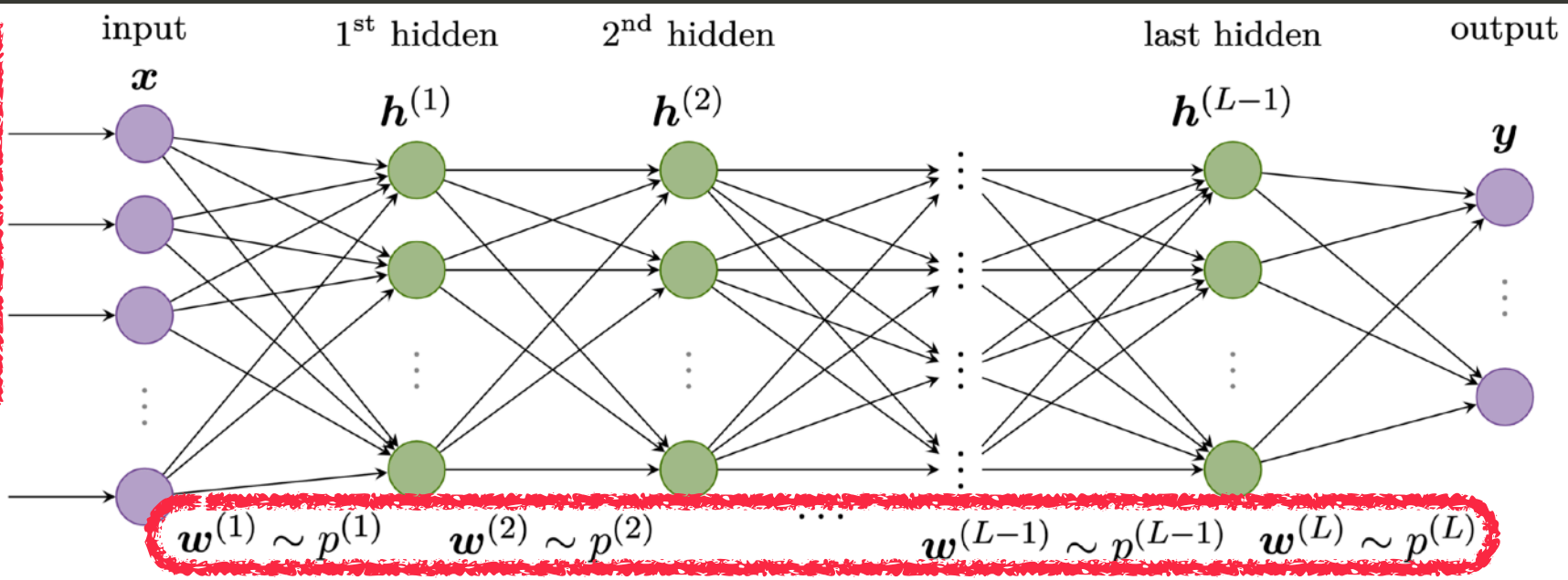
LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...

Teaching

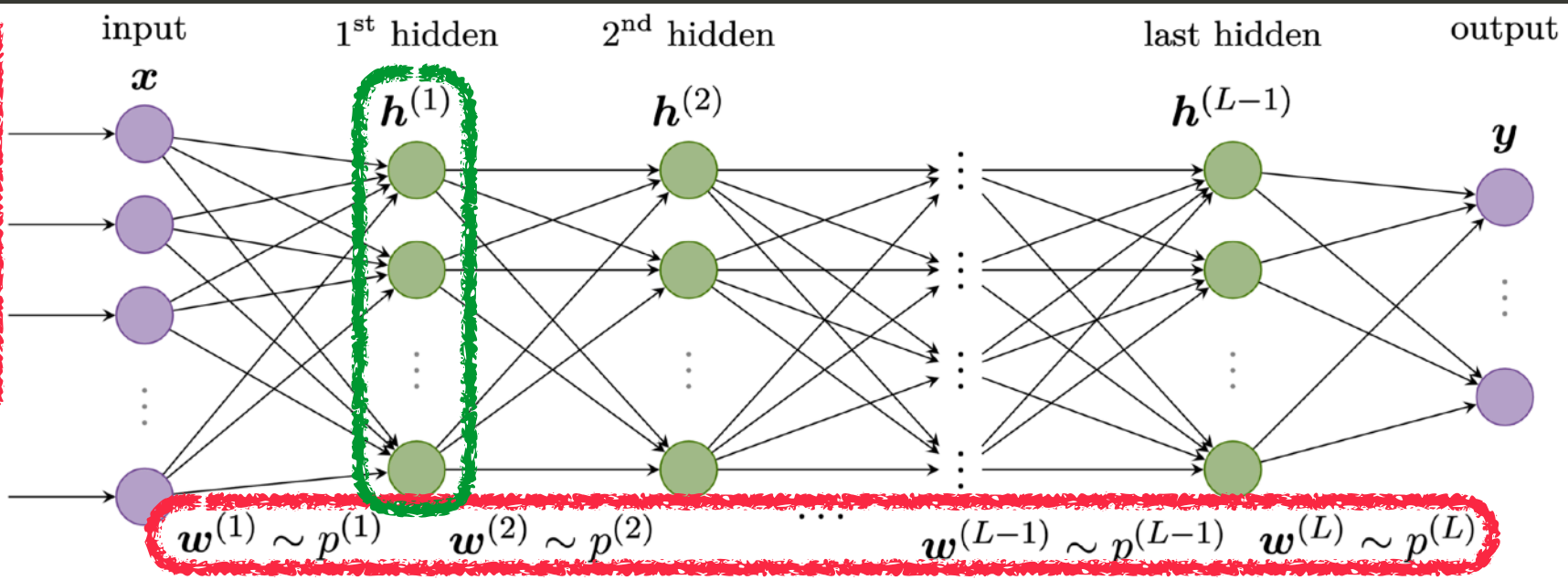
- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



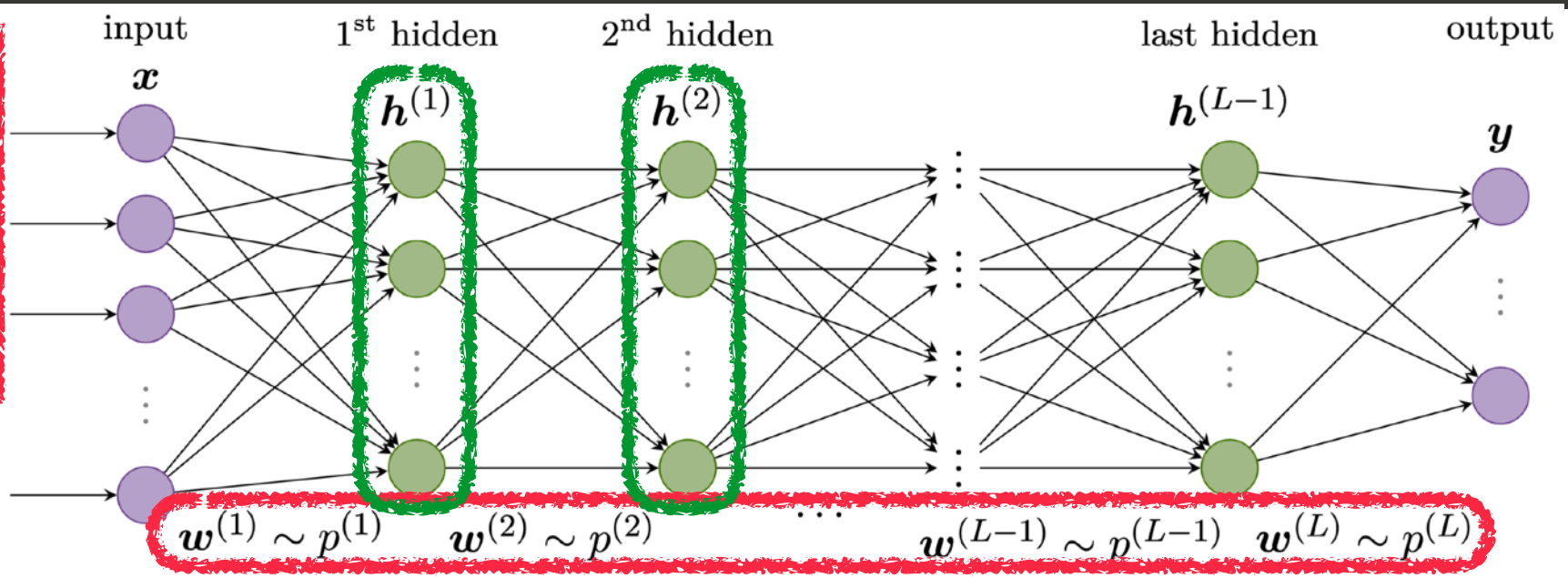
- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



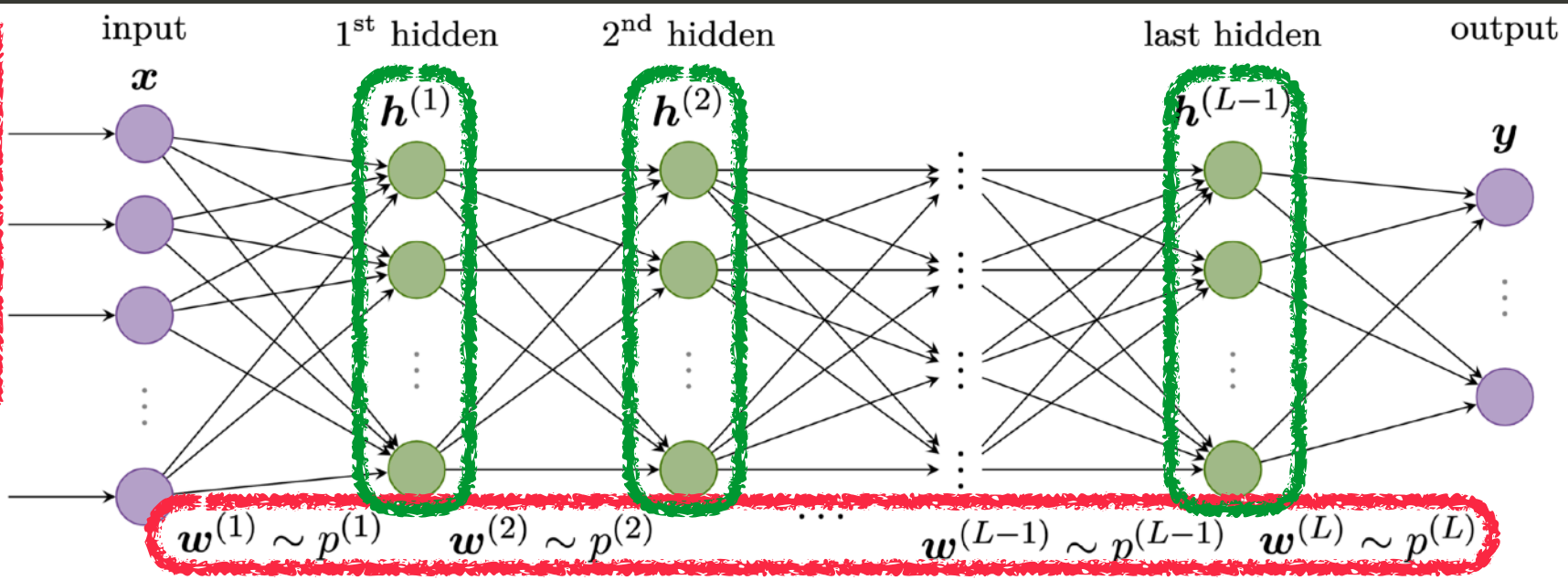
- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



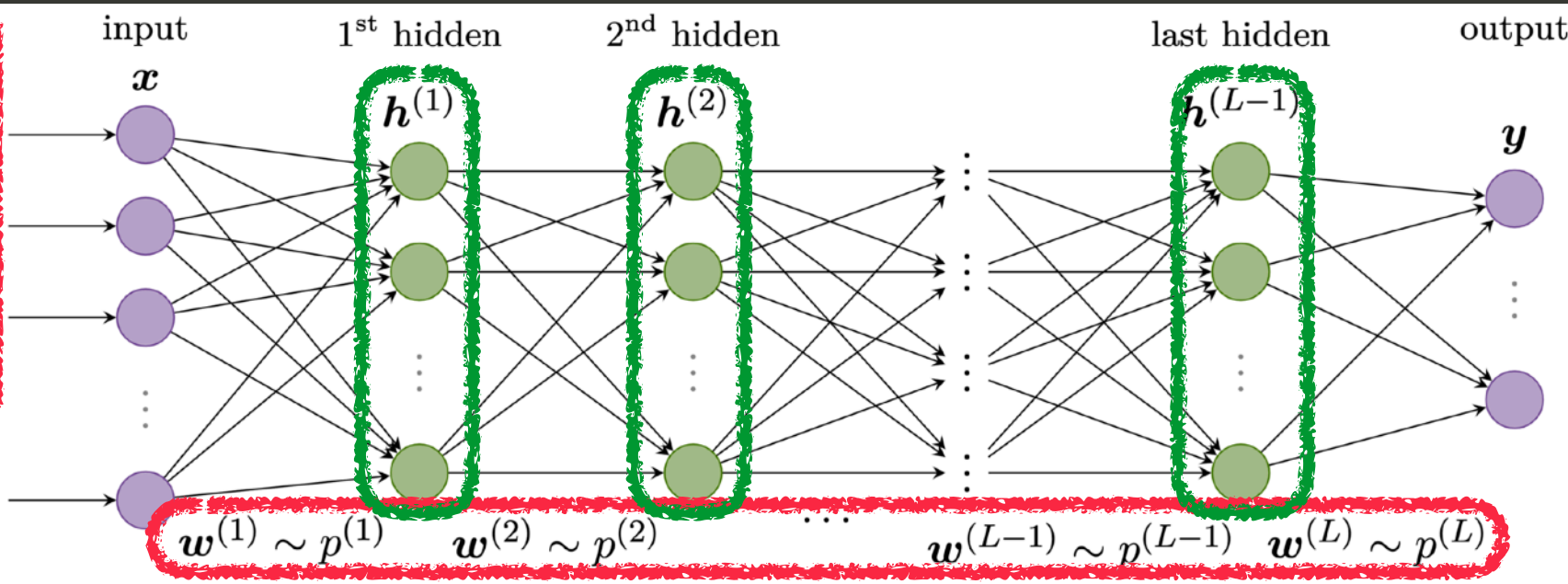
- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?

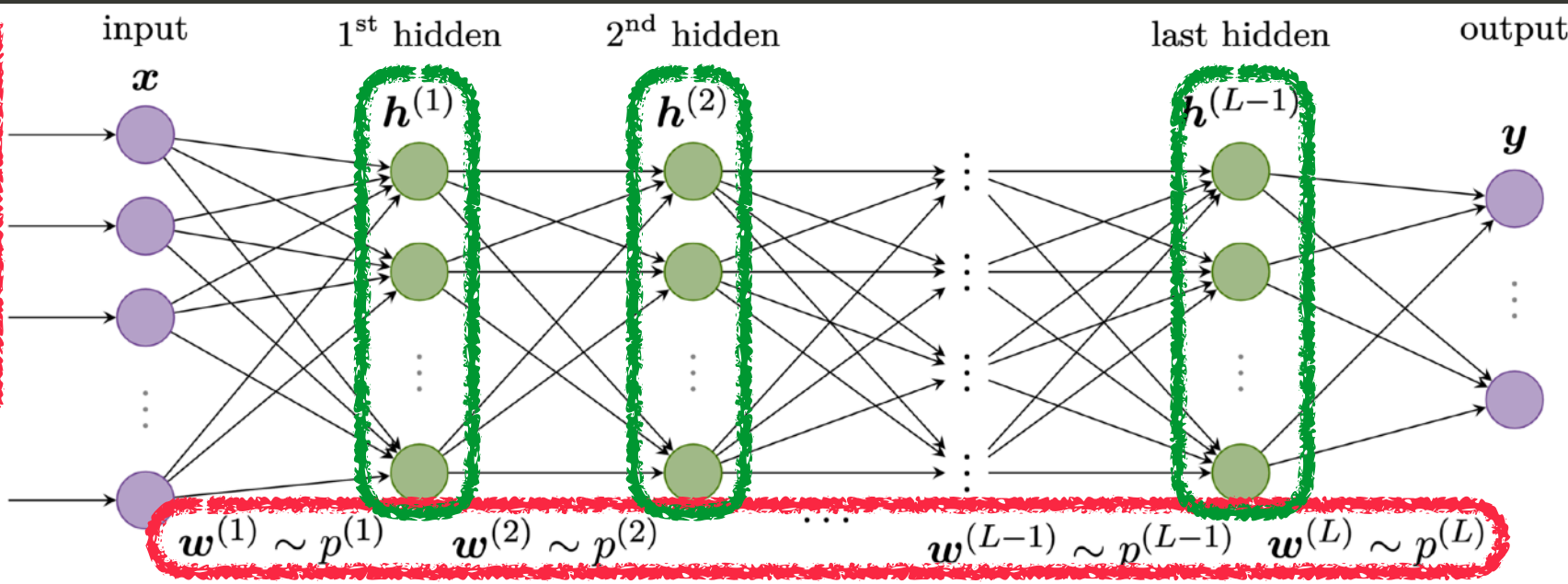


1 **When the network's width tends to infinity?**

Central limit theorem, Gaussian process

Edge of Chaos, Neural tangent kernel

- What prior distribution is induced on the **units** by a prior on the **weights**?
- Impact of width?
- Impact of depth?



1 **When the network's width tends to infinity?**
Central limit theorem, Gaussian process
Edge of Chaos, Neural tangent kernel

2 **When the network's width is kept fixed?**
Heavy-tails, sub-Weibull distribution, Weibull-tails
Sparsity inducing

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!

Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!

Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models

RM Neal

Journal of computational and graphical statistics 9 (2), 249-265

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!

Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!

Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

Slice sampling
RM Neal
Annals of statistics, 705-741

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!
Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

Probabilistic inference using Markov chain Monte Carlo methods
RM Neal
Department of Computer Science, University of Toronto

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

Slice sampling
RM Neal
Annals of statistics, 705-741

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!
Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

Probabilistic inference using Markov chain Monte Carlo methods
RM Neal
Department of Computer Science, University of Toronto

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

Annealed importance sampling
RM Neal
Statistics and computing 11, 125-139

Slice sampling
RM Neal
Annals of statistics, 705-741

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!
Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

Probabilistic inference using Markov chain Monte Carlo methods
RM Neal
Department of Computer Science, University of Toronto

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

Annealed importance sampling
RM Neal
Statistics and computing 11, 125-139

Slice sampling
RM Neal
Annals of statistics, 705-741

Bayesian learning for neural networks
RM Neal
Springer Science & Business Media

1 When the network's width tends to infinity

Radford Neal

U Toronto *** GOOGLE SCHOLAR CAN GIVE WRONG PUBLICATION DATE/REFERENCE - LOOK AT ACTUAL PAPER/BOOK!
Verified email at utstat.toronto.edu - Homepage

Bayesian inference Monte Carlo methods Information theory Machine learning Neural networks

Markov chain sampling methods for Dirichlet process mixture models
RM Neal
Journal of computational and graphical statistics 9 (2), 249-265

MCMC Using Hamiltonian Dynamics
R Neal
Handbook of Markov Chain Monte Carlo, 113-162

Slice sampling
RM Neal
Annals of statistics, 705-741

Probabilistic inference using Markov chain Monte Carlo methods
RM Neal
Department of Computer Science, University of Toronto

Annealed importance sampling
RM Neal
Statistics and computing 11, 125-139

Bayesian learning for neural networks
RM Neal
Springer Science & Business Media

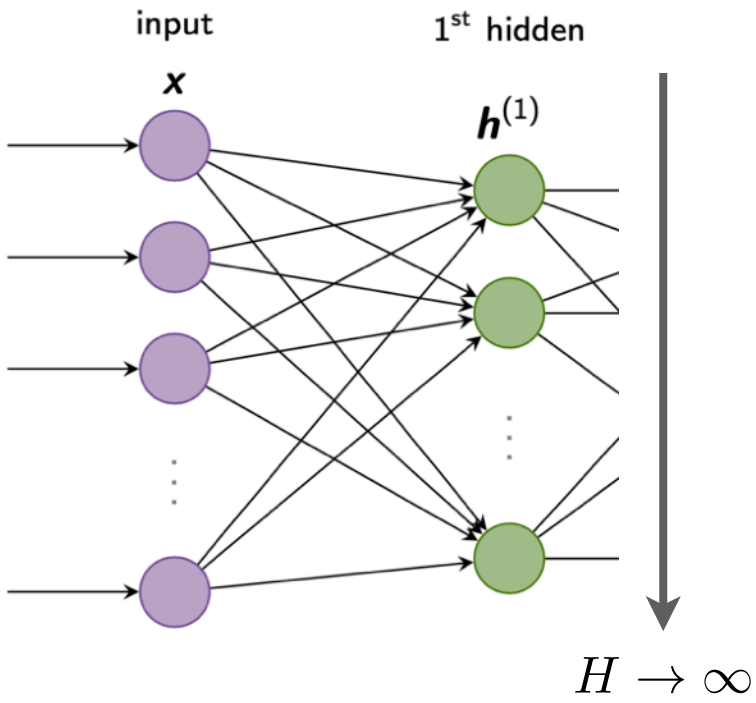
1 When the network's width tends to infinity

Bayesian learning for neural networks
RM Neal
Springer Science & Business Media

1 When the network's width tends to infinity

Bayesian learning for neural networks
RM Neal
Springer Science & Business Media

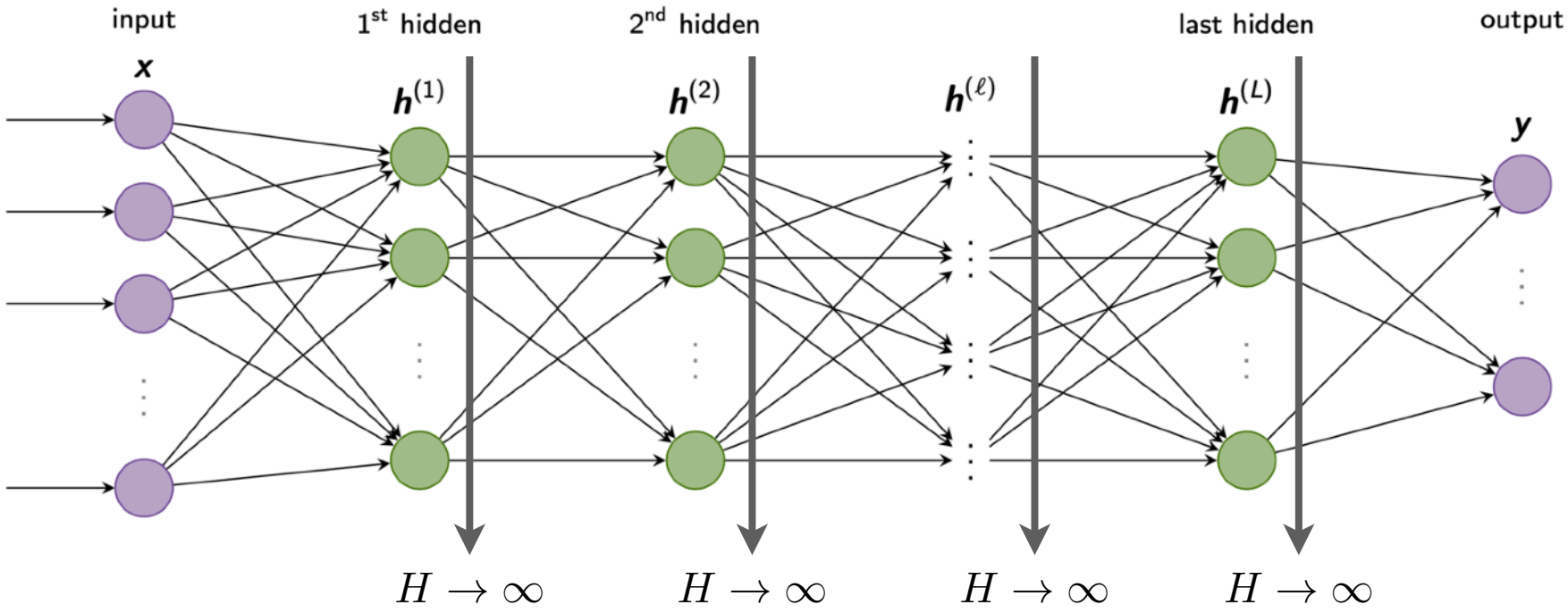
A one-hidden-layer neural network, whose width goes to infinity, and which has a Gaussian prior on all the parameters, converges to a Gaussian process with a well-defined kernel (Neal, 1996).



1 When the network's width tends to infinity

Gaussian Process Behaviour in Wide Deep Neural Networks
AGG Matthews, J Hron, M Rowland, RE Turner, Z Ghahramani
International Conference on Learning Representations (ICLR)

A deep neural network, where all layers widths go to infinity, and which has a Gaussian prior on all the parameters, converges to a Gaussian process with a well-defined kernel (Matthews et al., 2018).



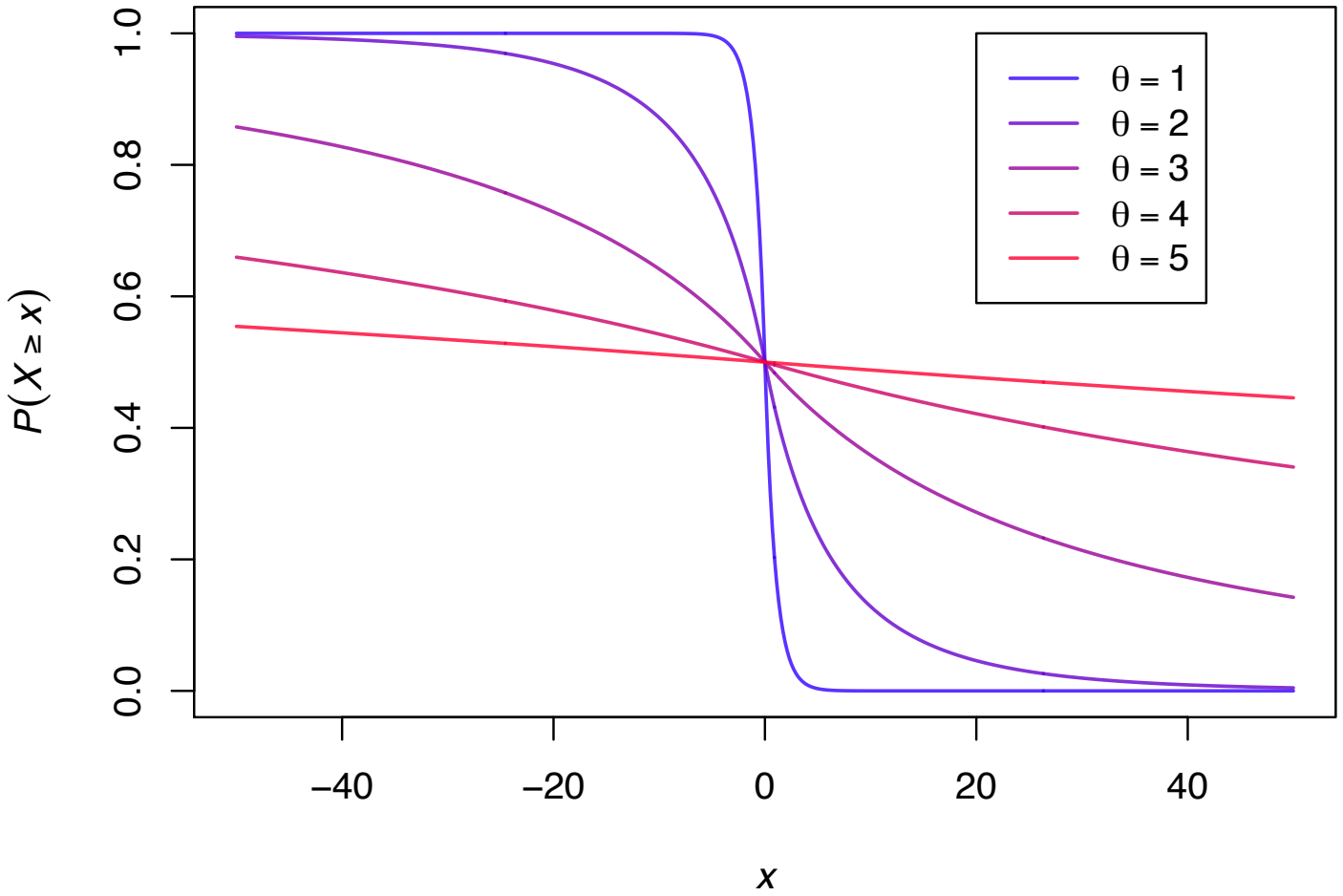
2 When the network's width is kept fixed?

2 When the network's width is kept fixed?

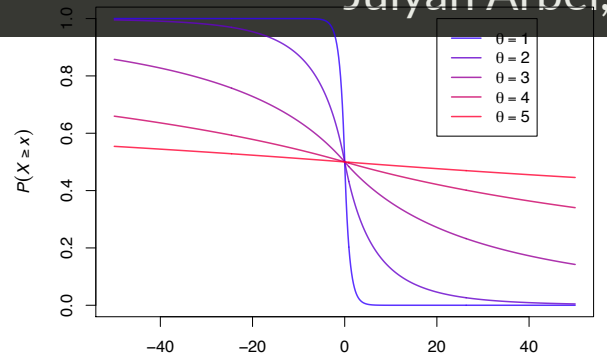
Sub-Weibull distributions

2 When the network's width is kept fixed?

Sub-Weibull distributions

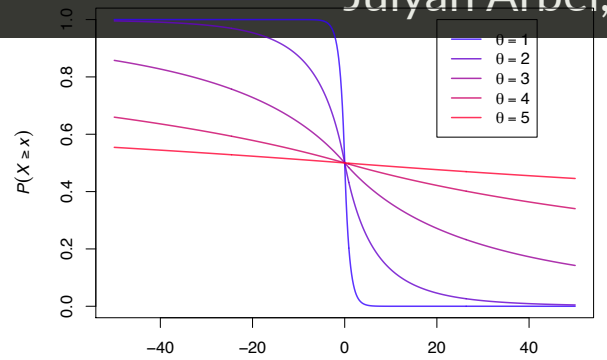


2 When the network's width is kept fixed?



Theorem. The rv X is Sub-Weibull iff:

2 When the network's width is kept fixed?

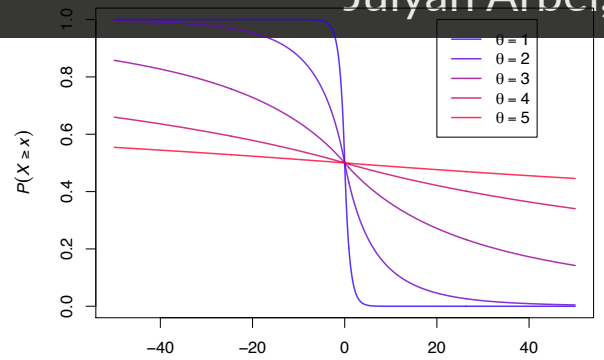


Theorem. The rv X is Sub-Weibull iff:

1. The tails of X satisfy

$$\exists K_1 > 0 \quad \text{such that} \quad \mathbb{P}(|X| \geq x) \leq 2\exp\left(-(x/K_1)^{1/\theta}\right) \quad \text{for all } x \geq 0.$$

2 When the network's width is kept fixed?



Theorem. The rv X is Sub-Weibull iff:

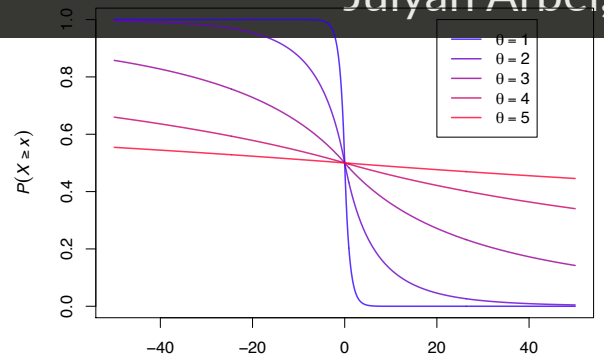
1. The tails of X satisfy

$$\exists K_1 > 0 \quad \text{such that} \quad \mathbb{P}(|X| \geq x) \leq 2\exp\left(-(x/K_1)^{1/\theta}\right) \quad \text{for all } x \geq 0.$$

2. The moments of X satisfy

$$\exists K_2 > 0 \quad \text{such that} \quad \|X\|_k \leq K_2 k^{\theta} \quad \text{for all } k \geq 1.$$

2 When the network's width is kept fixed?



Theorem. The rv X is Sub-Weibull iff:

1. The tails of X satisfy

$$\exists K_1 > 0 \quad \text{such that} \quad \mathbb{P}(|X| \geq x) \leq 2\exp\left(-(x/K_1)^{1/\theta}\right) \quad \text{for all } x \geq 0.$$

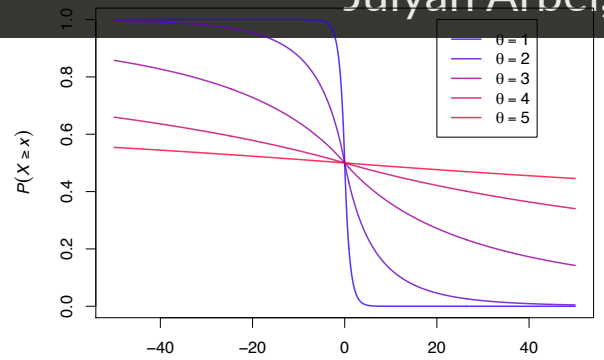
2. The moments of X satisfy

$$\exists K_2 > 0 \quad \text{such that} \quad \|X\|_k \leq K_2 k^{\theta} \quad \text{for all } k \geq 1.$$

3. The MGF of $|X|^{1/\theta}$ satisfies

$$\exists K_3 > 0 \quad \text{such that} \quad \mathbb{E}\left[\exp\left((\lambda|X|)^{1/\theta}\right)\right] \leq \exp\left((\lambda K_3)^{1/\theta}\right)$$

2 When the network's width is kept fixed?



Theorem. The rv X is Sub-Weibull iff:

1. The tails of X satisfy

$$\exists K_1 > 0 \quad \text{such that} \quad \mathbb{P}(|X| \geq x) \leq 2\exp\left(-(x/K_1)^{1/\theta}\right) \quad \text{for all } x \geq 0.$$

2. The moments of X satisfy

$$\exists K_2 > 0 \quad \text{such that} \quad \|X\|_k \leq K_2 k^{\theta} \quad \text{for all } k \geq 1.$$

3. The MGF of $|X|^{1/\theta}$ satisfies

$$\exists K_3 > 0 \quad \text{such that} \quad \mathbb{E}\left[\exp\left((\lambda|X|)^{1/\theta}\right)\right] \leq \exp\left((\lambda K_3)^{1/\theta}\right)$$

for all λ such that $0 < \lambda \leq 1/K_3$.

4. The MGF of $|X|^{1/\theta}$ is bounded at some point, namely

$$\exists K_4 > 0 \quad \text{such that} \quad \mathbb{E}\left[\exp\left((|X|/K_4)^{1/\theta}\right)\right] \leq 2.$$

2 When the network's width is kept fixed?

Assumptions.

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(0, \sigma^2)$$

(A2) Nonlinearity. ReLU-like with envelope property: $\exists\ c_1, c_2, d_2 \geq 0,\ d_1 > 0$ s.t.

$$|\phi(u)| \geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-,$$
$$|\phi(u)| \leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}.$$

2 When the network's width is kept fixed?

Assumptions.

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(0, \sigma^2)$$

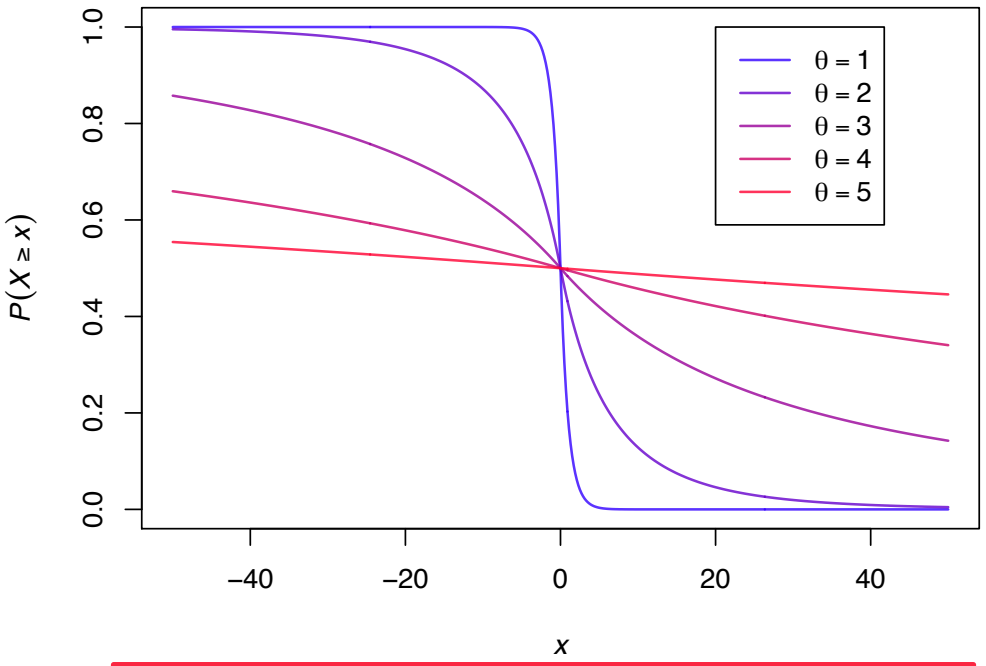
(A2) Nonlinearity. ReLU-like with envelope property: $\exists\ c_1, c_2, d_2 \geq 0,\ d_1 > 0$ s.t.

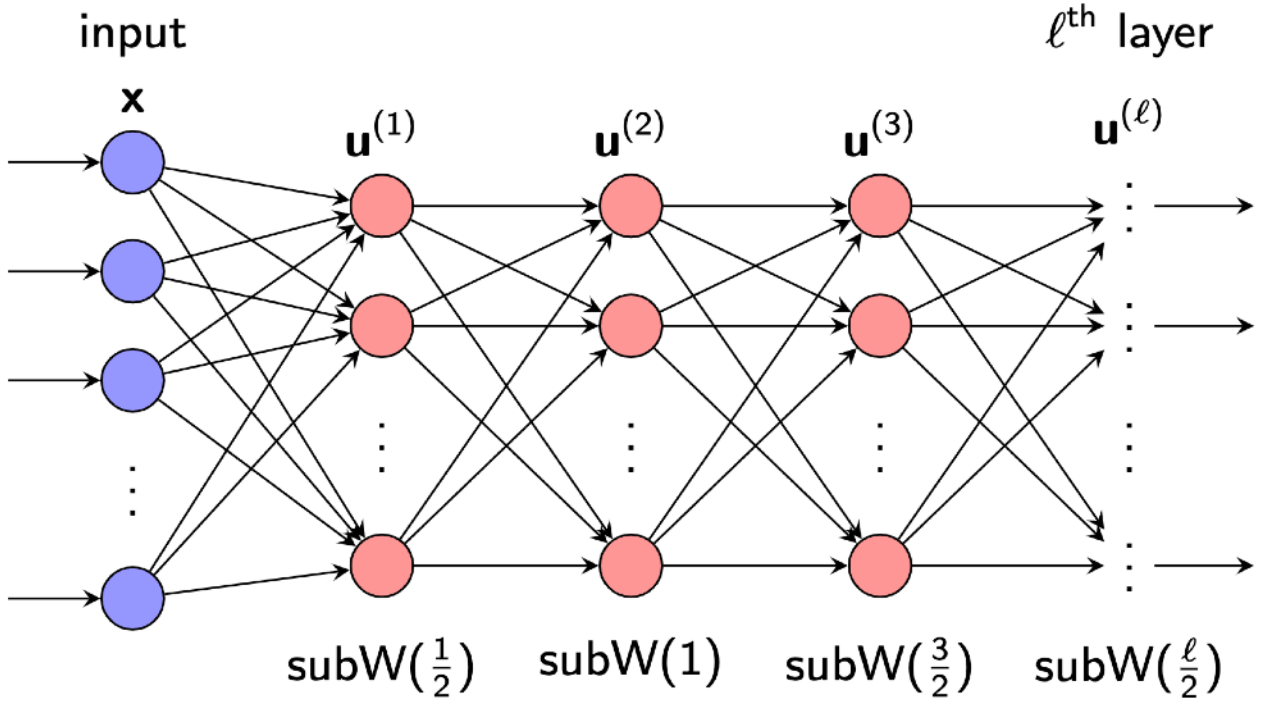
$$|\phi(u)| \geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-,$$
$$|\phi(u)| \leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}.$$

Theorem. sub-Weibull units.

Consider a Bayesian neural network with (A1) i.i.d. Gaussian priors on the weights and (A2) nonlinearity satisfying envelope property.
Then conditional on input $\mathbf{x}$, the marginal prior distribution of a unit $u^{(\ell)}$ of ℓ-th hidden layer is sub-Weibull with tail parameter $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim \text{subW}(\ell/2)$

2 When the network's width is kept fixed?





Theorem.

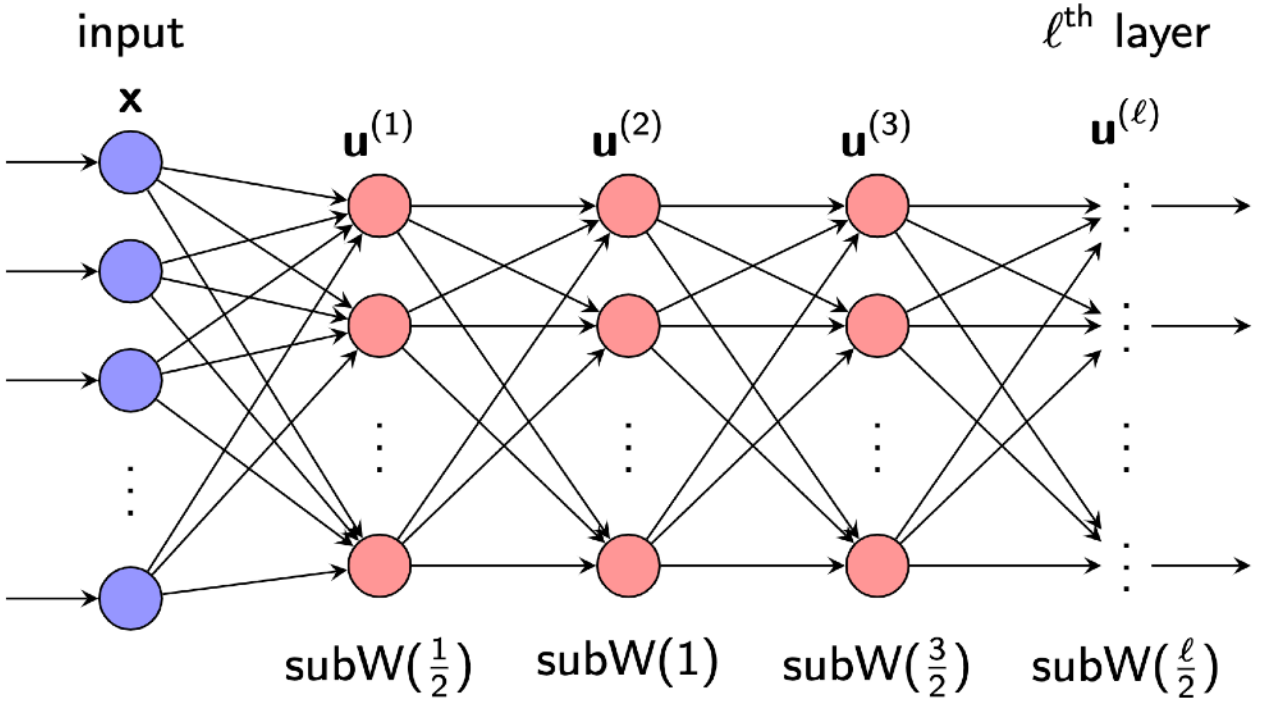
Consider a Bayesian neural network with (A1) i.i.d. Gaussian priors on the weights and (A2) nonlinearity satisfying envelope property.

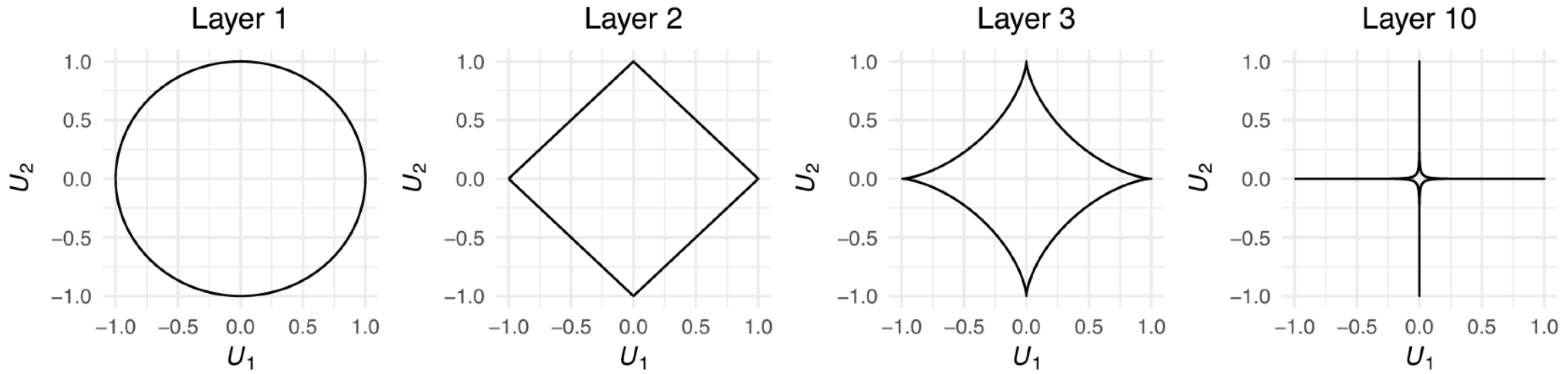
Then conditional on input $\mathbf{x}$, the marginal prior distribution of a unit $u^{(\ell)}$ of ℓ-th hidden layer is sub-Weibull with tail parameter $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim \text{subW}(\ell/2)$

2 When the network's width is kept fixed?

Weight distribution $\pi(w) \approx e^{-w^2}$ $\Rightarrow$ ℓ-th layer unit distribution $\pi^{(\ell)}(u) \approx e^{-u^{2/\ell}}$

Layer	Penalty on $\mathbf{W}$	Penalty on $\mathbf{U}$	
1	$\|\mathbf{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\mathbf{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\mathbf{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

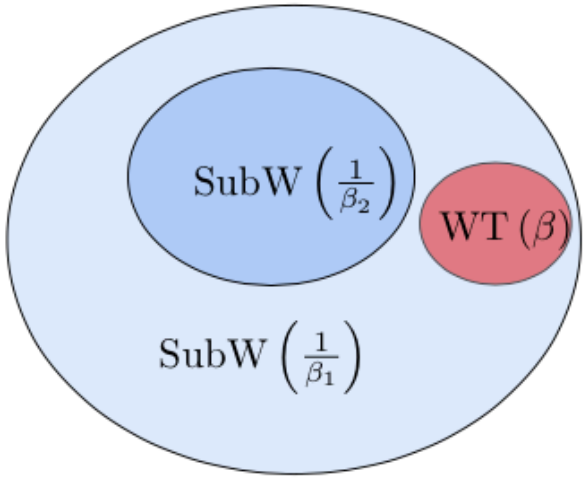




Sub-Weibull
(upper bound)

$$\overline{F}(x) \leq a\mathrm{e}^{-bx^{1/\theta}}, \quad \text{for } x > 0 \text{ and some } a, b, \theta > 0.$$

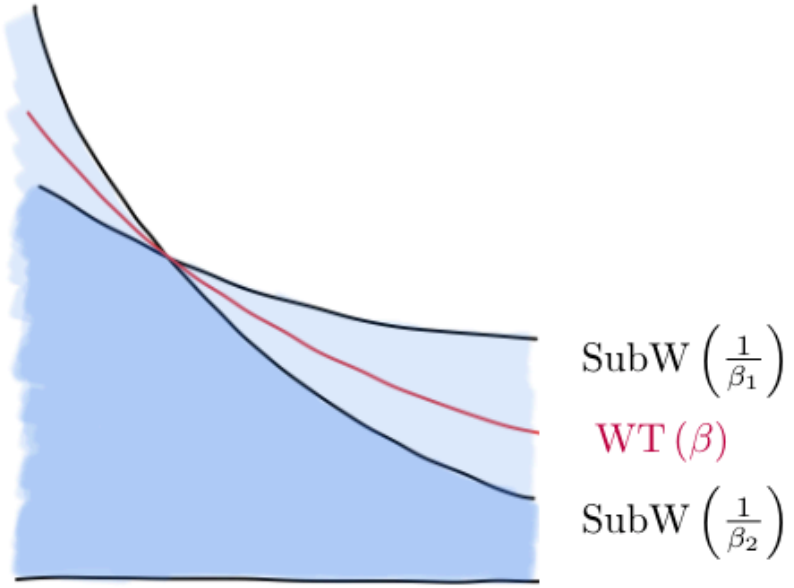
$$0 < \beta_2 < \beta < \beta_1$$



Mariia Vladimirova, Stéphane Girard, Hien D Nguyen, and Julyan Arbel.
Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. Stat, 2020.

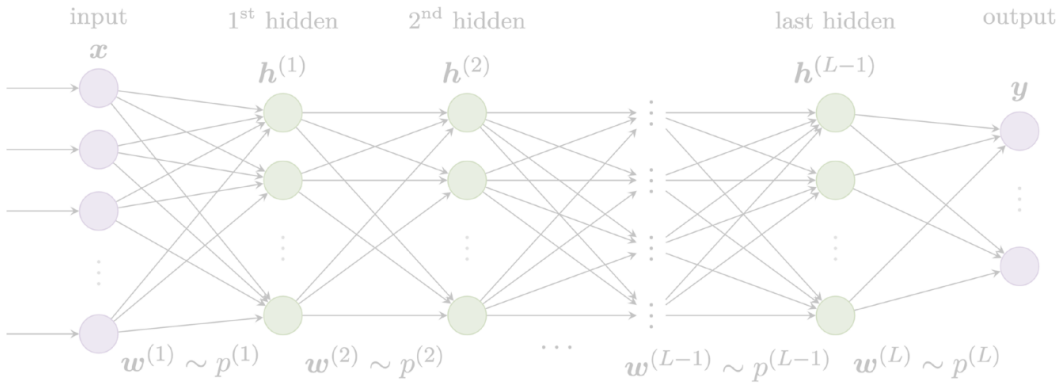
Generalized Weibull-tail
(precise description)

$$\mathrm{e}^{-x^{\beta} l_1(x)} \leq \overline{F}(x) \leq \mathrm{e}^{-x^{\beta} l_2(x)}, \quad \text{for } x > 0 \text{ and some } \beta > 0.$$

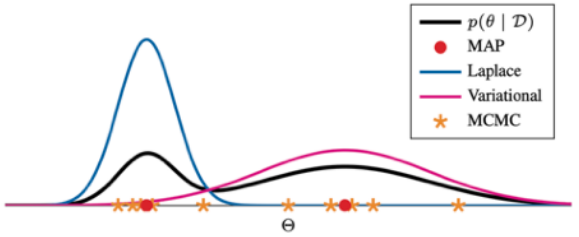


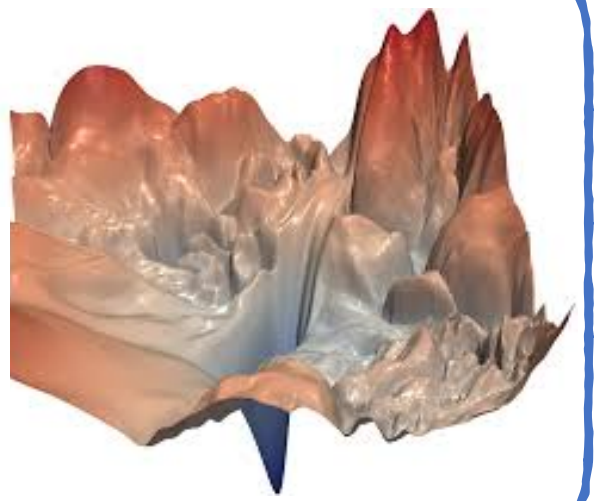
Mariia Vladimirova, Julyan Arbel, and Stéphane Girard.
Bayesian neural network unit priors and generalized Weibull-tail property. ACML, 2021.

Priors



Posteriors





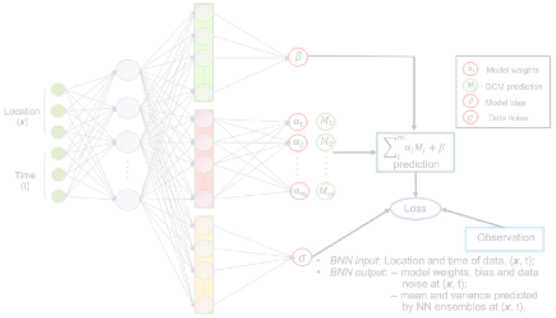
Applications

Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

Correct answer wrong, it is "sulfuric acid"

Bing Chat Yes, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...



Teaching

Posterior computations

Armed with a prior for our BNN, we still need to work out the posterior.

The BNN posterior is highly intractable.

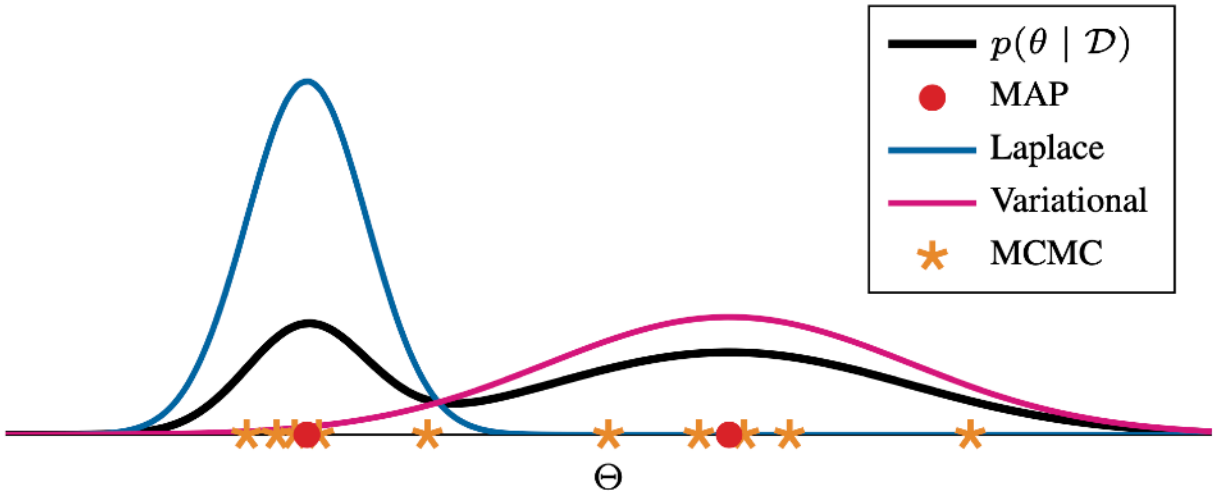
Can't calculate? Then approximate!

1. Approximate inference algorithm
2. Implementation of that algorithm

$$\underbrace{p(\boldsymbol{\theta} \mid \mathcal{D})}_{\text{posterior dist.}} = \frac{\overbrace{p(\mathcal{D} \mid \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta})}^{\text{prior dist.}}}{\underbrace{\int p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}_{\text{marginal likelihood}}}$$

1. Approximate inference algorithm

Different flavors of BDL methods for approximating the posterior



- Maximum a posteriori (MAP): point estimate
- MCMC/Hamiltonian MC (Neal, 1996)
- MC Dropout random masking of neurons at test time (Gal and Ghahramani, ICML, 2016)
- Variational inference and Laplace approximation: Gaussian approximations
- Ensemble methods use MAP estimates as their samples.
- Partially Stochastic Bayesian Neural Networks (Sharma et al. AISTATS, 2023)

Maximum a posteriori estimation (MAP)

Rely on SGD to find a maximum a posteriori estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n}, y_{1:n})$$

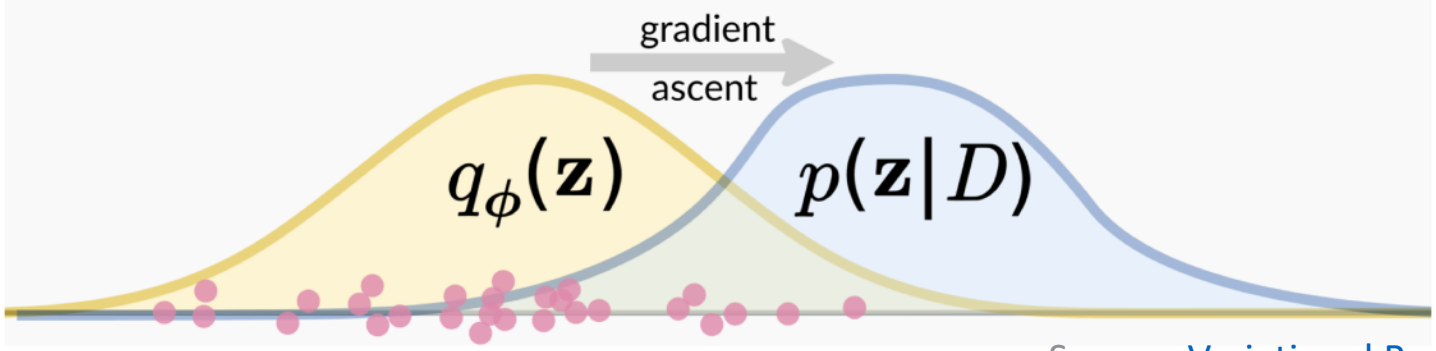
$$= \arg\min_{\boldsymbol{\theta}} - \log p(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n}, y_{1:n})$$

$$= \arg\min_{\boldsymbol{\theta}} - \log p(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})$$

Variational inference (VI)

Goal is to approximate the posterior distribution by specifying an approximate family $q_\lambda(\cdot)$
Turns sampling into optimization, minimizing the evidence lower bound (ELBO)

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n}, y_{1:n}) = \frac{p(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \approx q(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \triangleq q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$$



Source: Variational Bayesian phylogenetic inference

Laplace approximation (LA)

Application of the Laplace method (P.S. Laplace, 1774) to approximate posterior distributions by a Gaussian centered at the posterior mode $\hat{\boldsymbol{\theta}}$, using the curvature (Hessian) of the log-posterior $\psi(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta} \,|\, x_{1:n}, y_{1:n})$.
Take the 2nd-order Taylor expansion of $\psi(\boldsymbol{\theta})$ around the MAP estimate $\hat{\boldsymbol{\theta}}$.

$$\psi(\boldsymbol{\theta}) \approx \psi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\Big[\ldots\Big] + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \Big[\nabla^2_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})\Big]_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

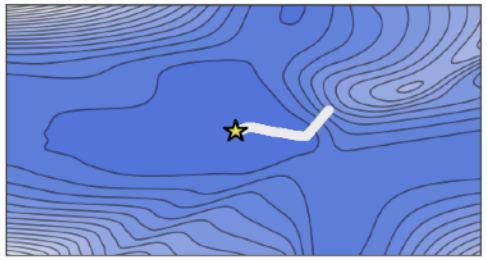
Compare this to the log-p.d.f.~of a multivariate Normal distribution.

$$\log \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) + \text{const.}$$

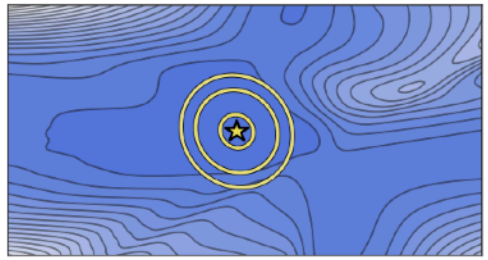
Laplace approximation: predictive posterior

How do we make predictions using the Laplace approximation?

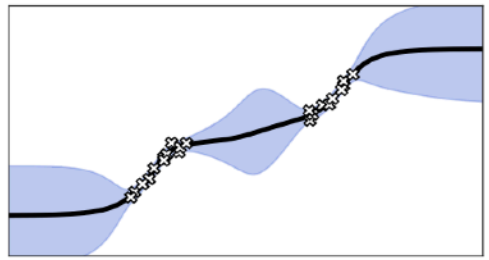
$$
\begin{aligned}
p(y^\star \mid \boldsymbol{x}^\star, \boldsymbol{x}_{1:n}, y_{1:n}) &= \int p(y^\star \mid \boldsymbol{x}^\star, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n}, y_{1:n})\, d\boldsymbol{\theta} \\
&\approx \int p(y^\star \mid \boldsymbol{x}^\star, \boldsymbol{\theta})\, q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim q} \Big[p(y^\star \mid \boldsymbol{x}^\star, \boldsymbol{\theta}) \Big]
\end{aligned}
$$



(a) Step 1: Find MAP



(b) Step 2: Fit approx.



(c) Step 3: Predict!

(Daxberger et al., 2021)

2. Implementation of the approx. algorithm

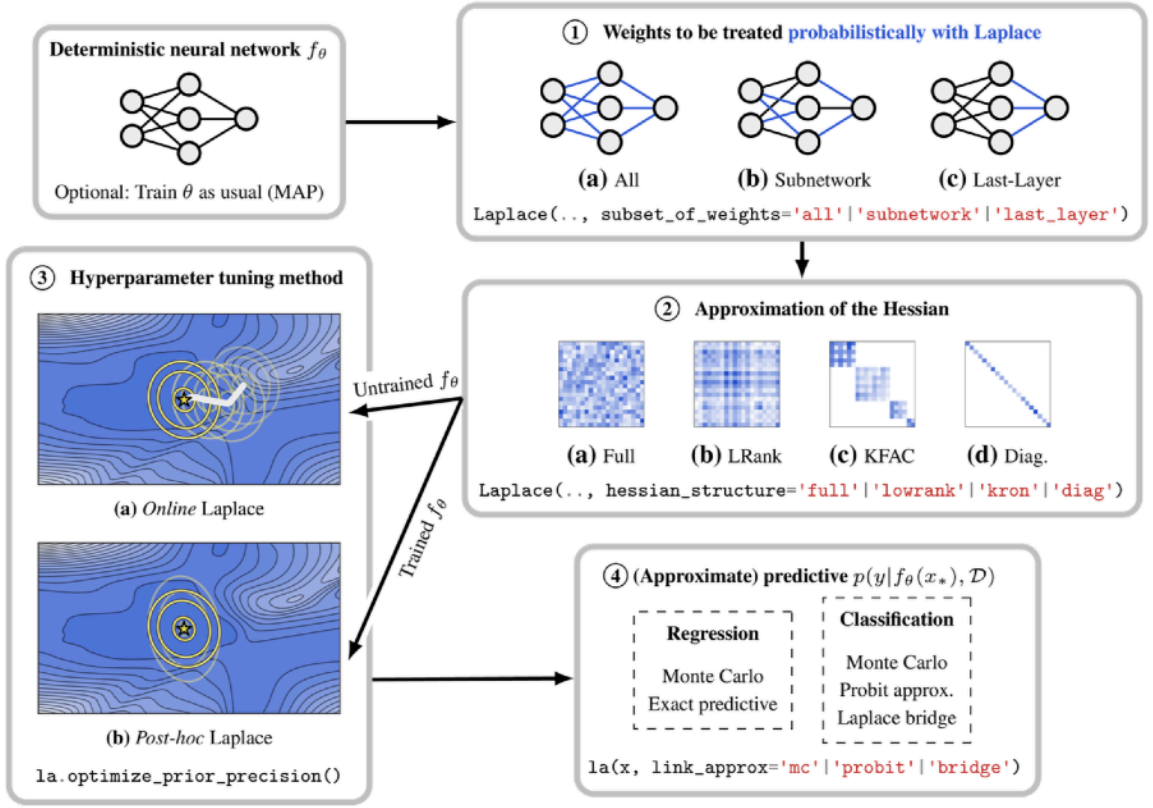
Unfortunately, there doesn't exist (yet) a unique/unified implementation framework. There exist quite a few software packages, libraries or probabilistic programming languages (PPLs) developed independently by different groups.

2. Implementation of the approx. algorithm

Unfortunately, there doesn't exist (yet) a unique/unified implementation framework. There exist quite a few software packages, libraries or probabilistic programming languages (PPLs) developed independently by different groups.

-> some algo. might only work with some normalization (layer norm vs. batch norm) or might not work with low-precision weights.
-> need for more user-friendly tools for layman practitioners

- Laplace inference: laplace-torch
  https://github.com/aleximmer/Laplace
- TyXe: Pyro-based BNNs for Pytorch users
  https://github.com/TyXe-BDL/TyXe
- BNN implementation using JAX
  https://neptune.ai/blog/bayesian-neural-networks-with-jax
- TensorFlow Probability
  https://www.tensorflow.org/probability



Theory about the BDL posterior distribution

Frequentist properties of the posterior for large sample size n

Nonparametric regression (Kohler and Langer, 2021; Castillo and Egels, 2024)

Multiclass classification (Bos and Schmidt-Hieber, 2022)

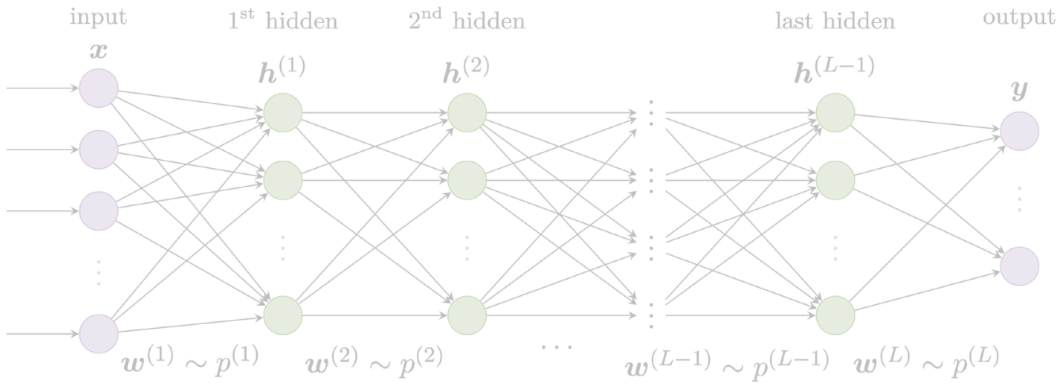
Study of spike-and-slab priors (Polson and Rockova, 2018; Jantre et al., 2024)

General priors (Lee and Lee, 2022; Kong and Kim, 2024)

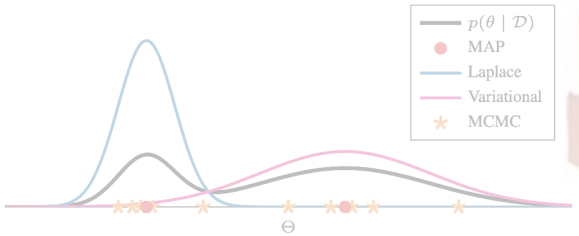
Deep Gaussian processes (Finocchio and Schmidt-Hieber, 2023; Bachoc and Lagnoux, 2025; Castillo and Randrianarisoa, 2024)

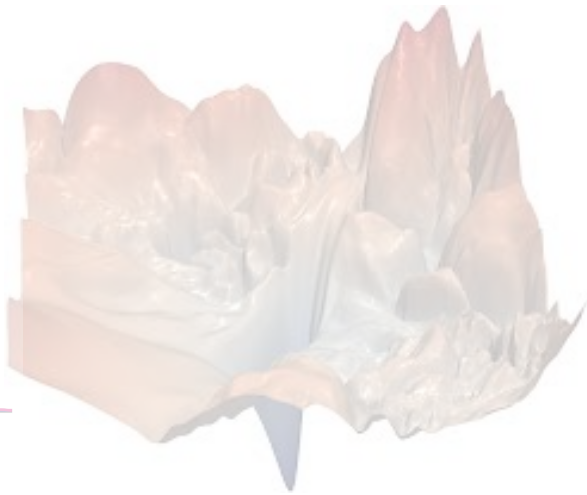
Variational approximations to BNN posteriors (Bhattacharya and Maiti (2021); Bhattacharya et al. (2020)

Priors



Posteriors





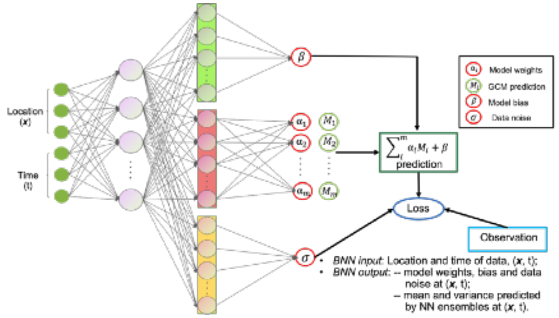
Applications

Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

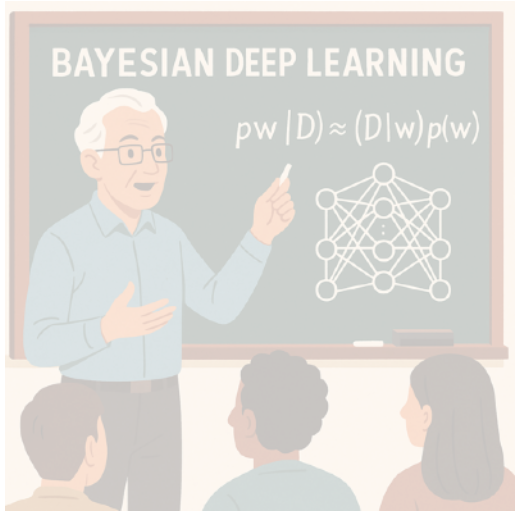
Correct answer wrong, it is "sulfuric acid"

Bing Chat Yes, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...



Teaching

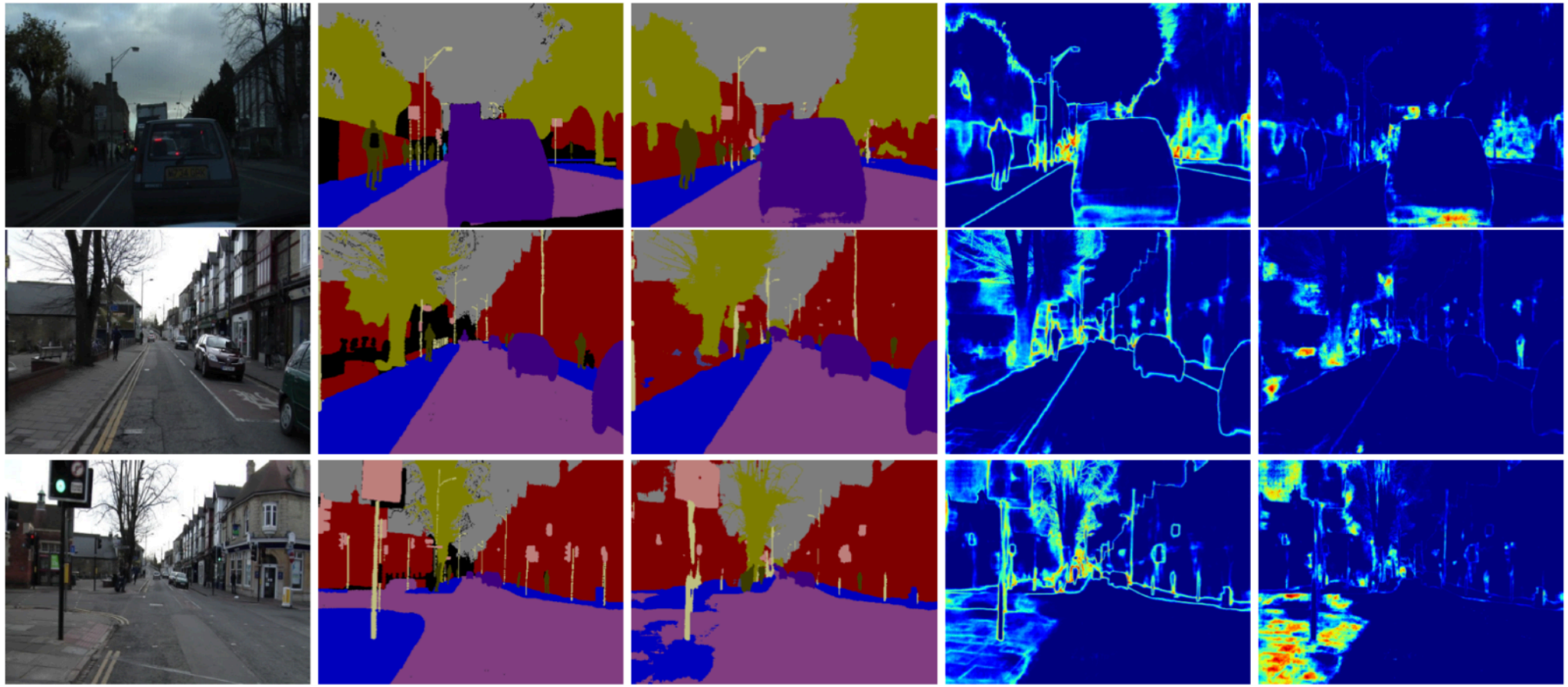


Why Bayesian deep learning matters?

BDL has shown potential in a range of critical application domains:

- healthcare (Peng et al., 2019; Abdar et al., 2021b)
- single-cell biology (Way and Greene, 2018),
- drug discovery (Gruver et al., 2021; Klarner et al., 2023),
- agriculture (Hernandez and Lopez, 2020),
- astrophysics (Soboczenski et al., 2018; Ferreira et al., 2020),
- nanotechnology (Leitherer et al., 2021),
- physics (Cranmer et al., 2021),
- climate science (Vandal et al., 2018; Luo et al., 2022),
- smart electricity grids (Yang et al., 2019),
- wearables (Manogaran et al., 2019; Zhou et al., 2020),
- robotics (Shi et al., 2021; Mur-Labadia et al., 2023),
- autonomous driving (McAllister et al., 2017).

Semantic segmentation

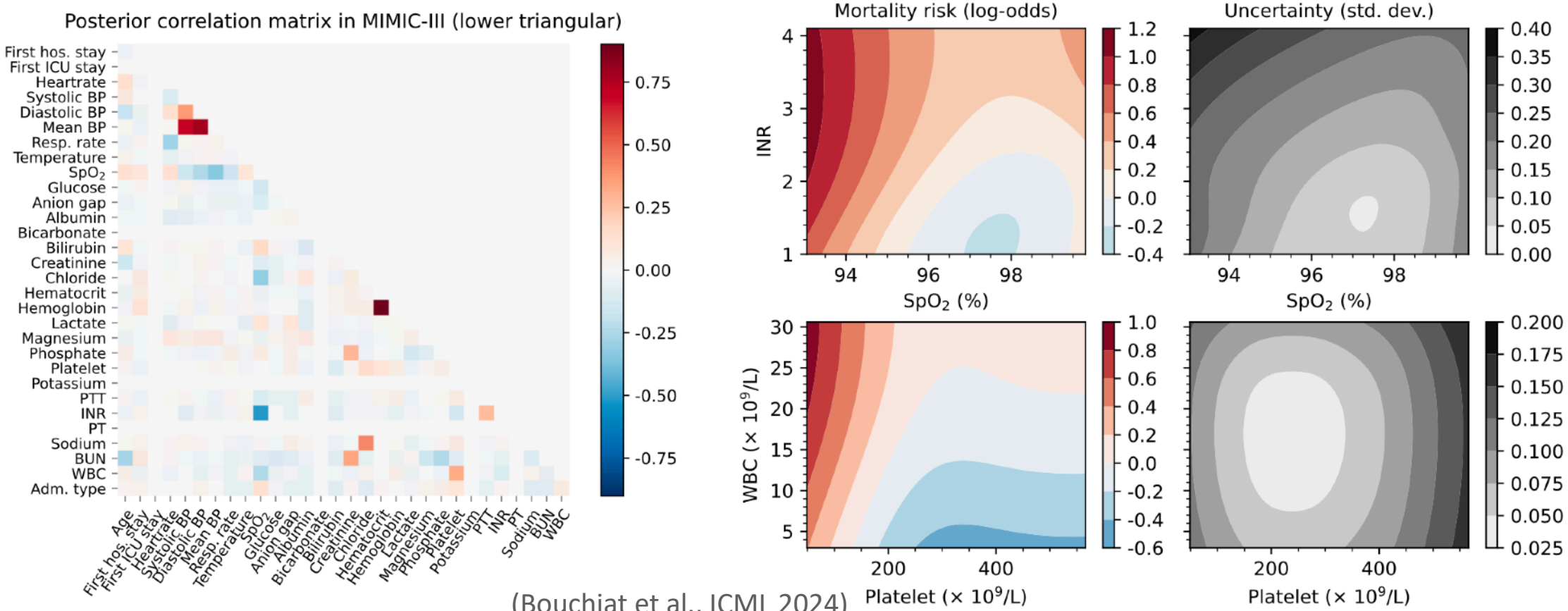


(a) Input Image | (b) Ground Truth | (c) Semantic Segmentation | (d) Aleatoric Uncertainty | (e) Epistemic Uncertainty

Kendall et al., 2017

Improvements in healthcare

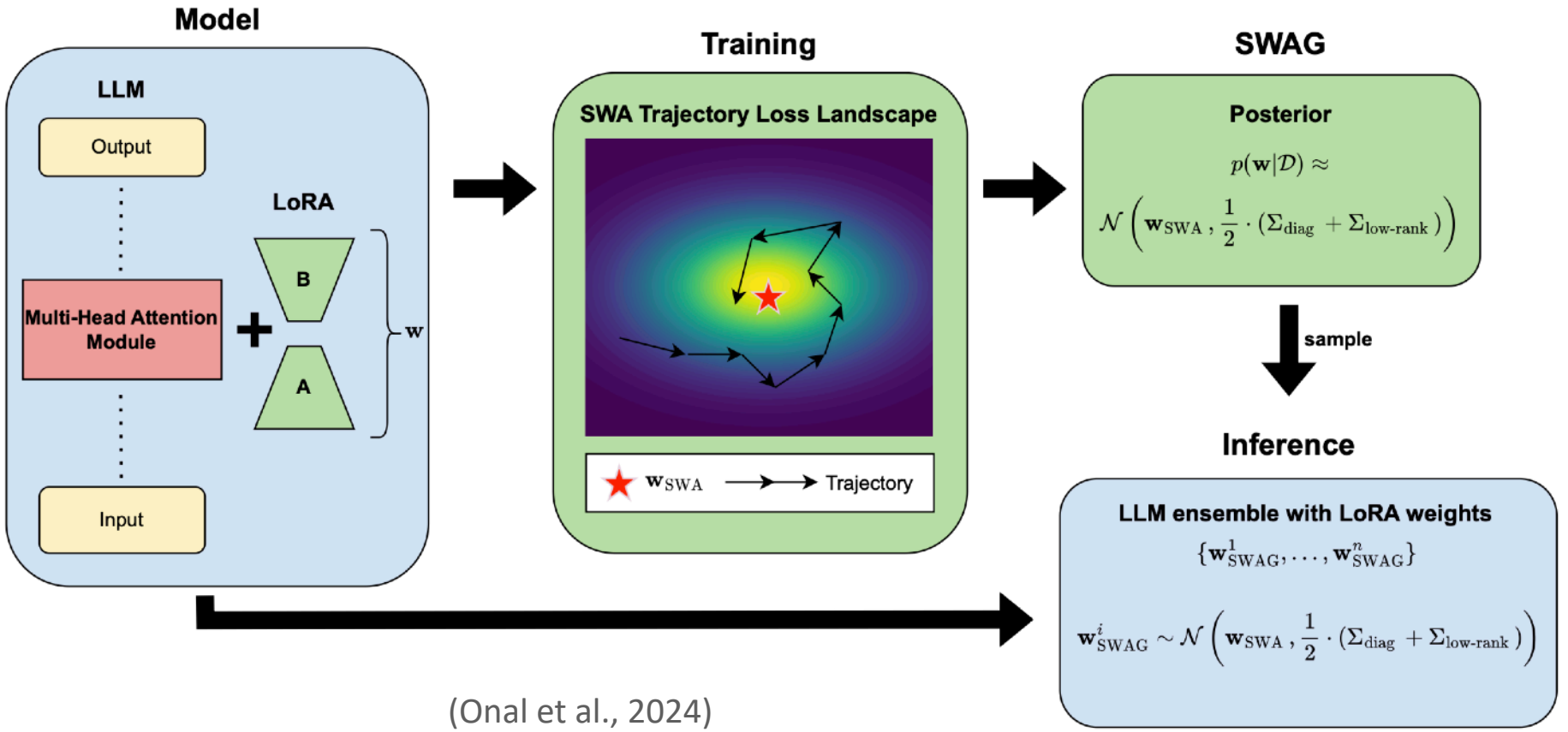
BDL can yield better predictions and uncertainty estimates, and lead to recommendations that are more interpretable for medical practitioners.



(Bouchiat et al., ICML 2024)

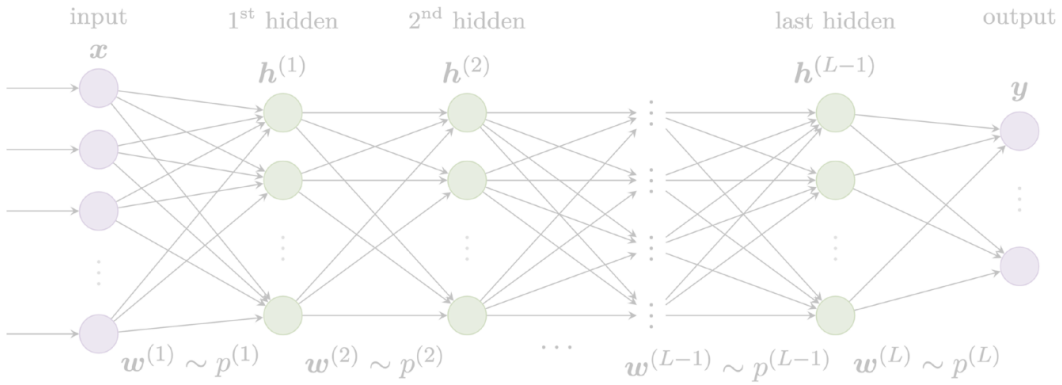
Large language models

Combination of Low-Rank Adaptation (LoRA) with Gaussian Stochastic Weight Averaging (SWAG) for LLMs fine-tuning on small datasets

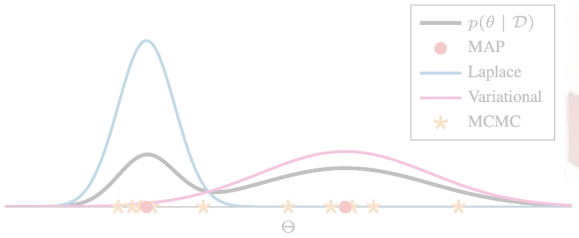


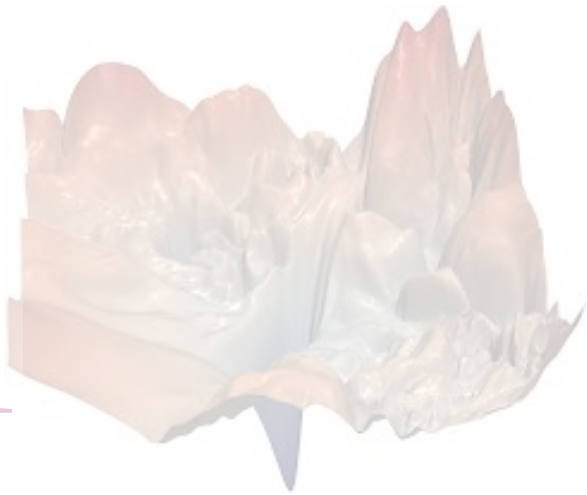
(Onal et al., 2024)

Priors



Posteriors





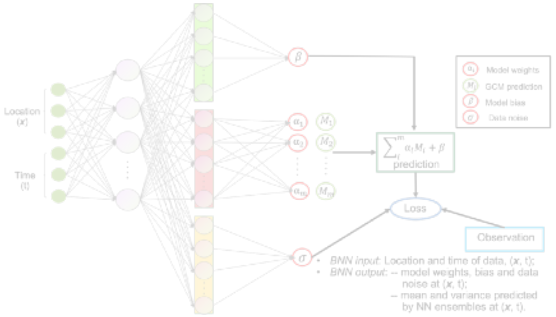
Applications

Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

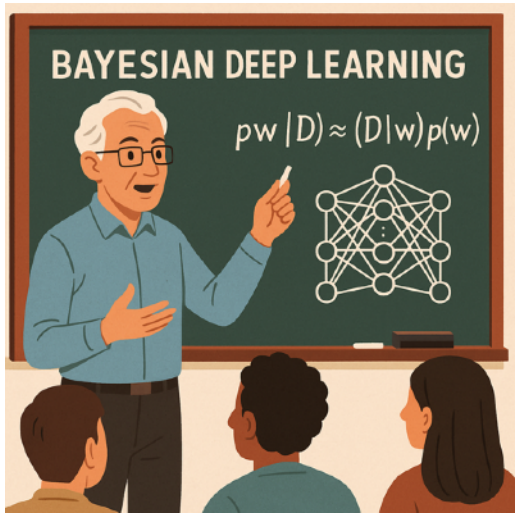
Correct answer wrong, it is "sulfuric acid"

Bing Chat Yes, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

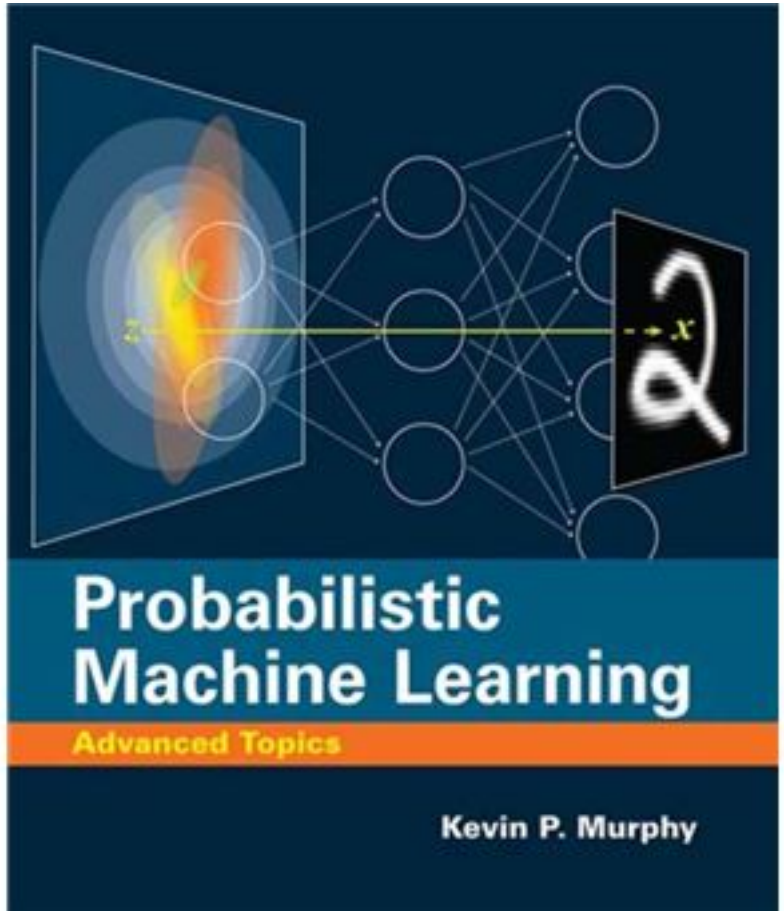
LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...



Teaching



Probabilistic Machine Learning by Kevin P. Murphy



A Primer on Bayesian Neural Networks: Review and Debates

Julyan Arbel[1], Konstantinos Pitas[1], Mariia Vladimirova[2], Vincent Fortuin[3]

[1] *Centre Inria de l'Université Grenoble Alpes, France*
[2] *Criteo AI Lab, Paris, France*
[3] *Helmholtz AI, Munich, Gremany*

Neural networks have achieved remarkable performance across various problem domains, but their widespread applicability is hindered by inherent limitations such as overconfidence in predictions, lack of interpretability, and vulnerability to adversarial attacks. To address these challenges, Bayesian neural networks (BNNs) have emerged as a compelling extension of conventional neural networks, integrating uncertainty estimation into their predictive capabilities.

This comprehensive primer presents a systematic introduction to the fundamental concepts of neural networks and Bayesian inference, elucidating their synergistic integration for the development of BNNs. The target audience comprises statisticians with a potential background in Bayesian methods but lacking deep learning expertise, as well as machine learners proficient in deep neural networks but with limited exposure to Bayesian statistics. We provide an overview of commonly employed priors, examining their impact on model behavior and performance. Additionally, we delve into the practical considerations associated with training and inference in BNNs.

Furthermore, we explore advanced topics within the realm of BNN research, acknowledging the existence of ongoing debates and controversies. By offering insights into cutting-edge developments, this primer not only equips researchers and practitioners with a solid foundation in BNNs, but also illuminates the potential applications of this dynamic field. As a valuable resource, it fosters an understanding of BNNs and their promising prospects, facilitating further advancements in the pursuit of knowledge and innovation.



Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI

Theodore Papamarkou [1] **Maria Skoularidou** [2] **Konstantina Palla** [3] **Laurence Aitchison** [4] **Julyan Arbel** [5]
David Dunson [6] **Maurizio Filippone** [7] **Vincent Fortuin** [8 9 10] **Philipp Hennig** [11] **José Miguel Hernández-Lobato** [12]
Aliaksandr Hubin [13 14] **Alexander Immer** [15] **Theofanis Karaletsos** [16] **Mohammad Emtiyaz Khan** [17]
Agustinus Kristiadi [18] **Yingzhen Li** [19] **Stephan Mandt** [20] **Christopher Nemeth** [21] **Michael A. Osborne** [22]
Tim G. J. Rudner [23] **David Rügamer** [10 24] **Yee Whye Teh** [25 26] **Max Welling** [27] **Andrew Gordon Wilson** [28]
Ruqi Zhang [29]

Abstract

In the current landscape of deep learning research, there is a predominant emphasis on achieving high predictive accuracy in supervised tasks involving large image and language datasets. However, a broader perspective reveals a multitude of overlooked metrics, tasks, and data types, such as uncertainty, active and continual learning, and scientific data, that demand attention. Bayesian deep learning (BDL) constitutes a promising avenue, offering advantages across these diverse settings. This paper posits that BDL can elevate the capabilities of deep learning. It revisits the strengths of BDL, acknowledges existing challenges, and highlights some exciting research avenues aimed at addressing these obstacles. Looking ahead, the discussion focuses on possible ways to combine large-scale foundation models with BDL to unlock their full potential.

Bayesian methods are more critical than ever in the age of large-scale AI in order to reliably assess uncertainties and incorporate existing knowledge in safety-critical decision-making algorithms.

ICML 2024 https://arxiv.org/abs/2402.00809

Bayesian deep learning textbook - 2026 🤞

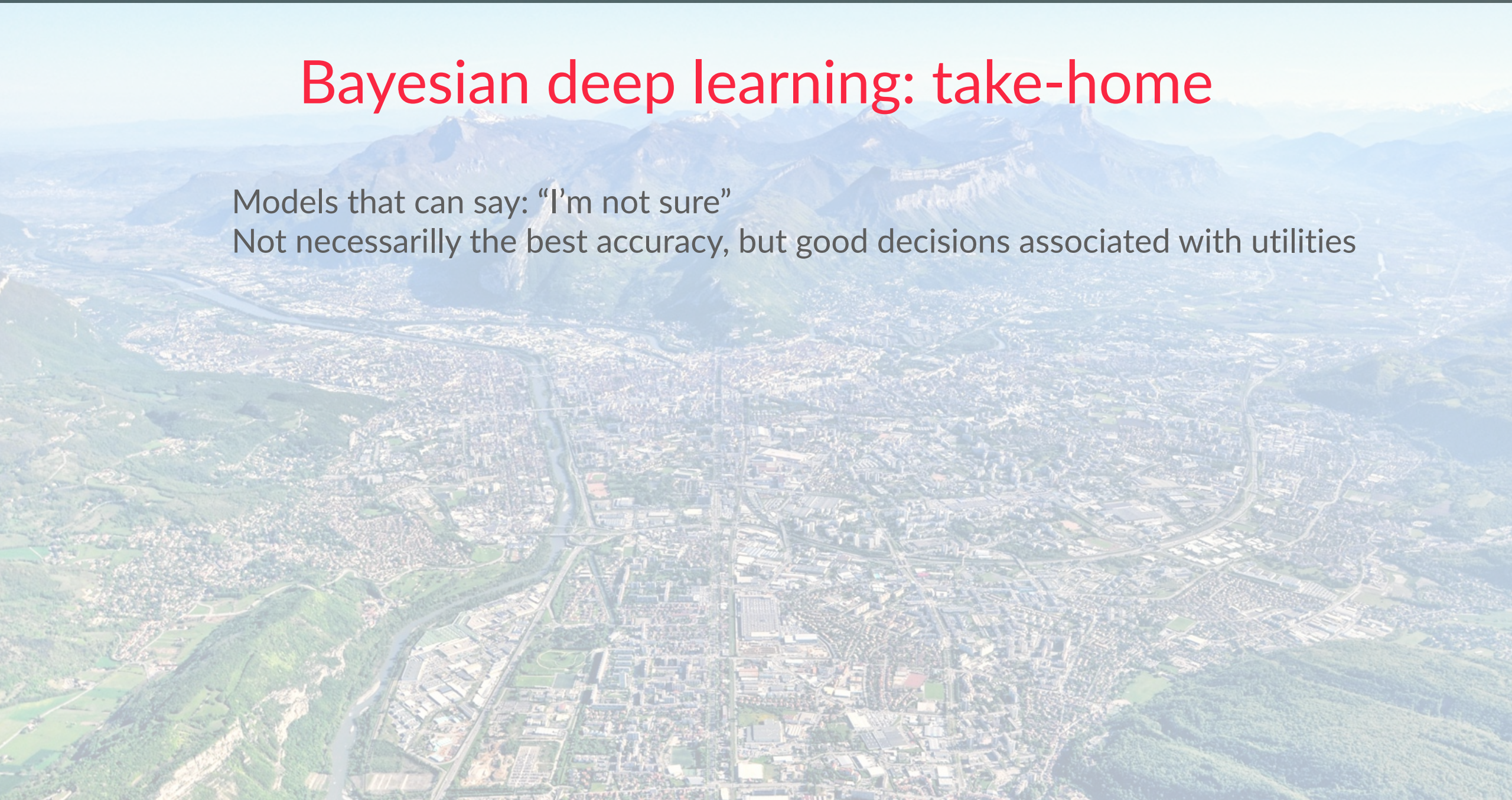
Large collaborative effort by a 'BDL consortium' of 100 researchers from various institutions

Part 1: Sampling methods
Part 2: Laplace approximations
Part 3: Variational inference
Part 4: Ensemble methods
Part 5: Kernel methods
Part 6: Priors
Part 7: Identifiability and symmetries
Part 8: Scalability with BDL and of BDL
Part 9: Applications
Part 10: Topical developments

Bayesian deep learning: take-home

Bayesian deep learning: take-home

Models that can say: “I’m not sure”
Not necessarilly the best accuracy, but good decisions associated with utilities



Bayesian deep learning: take-home

Models that can say: “I’m not sure”
Not necessarilly the best accuracy, but good decisions associated with utilities

Strengths of BDL

- Uncertainty Quantification
- Data Efficiency
- Adaptability, Interpretability

Weaknesses of BDL

- Scalability for posterior computation
- Prior specification

Bayesian deep learning: take-home

Models that can say: “I’m not sure”
Not necessarilly the best accuracy, but good decisions associated with utilities

Strengths of BDL

- Uncertainty Quantification
- Data Efficiency
- Adaptability, Interpretability

Weaknesses of BDL

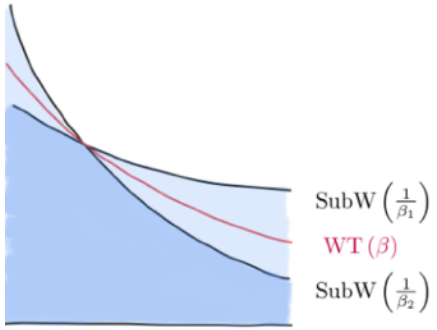
- Scalability for posterior computation
- Prior specification

Workflow

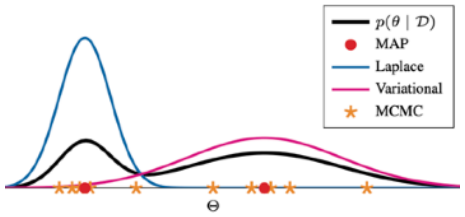
- Choose a prior
- Choose an approximation method
- Choose an implementation

References

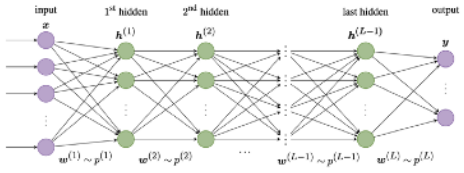
Thank you for your attention!



- Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. **Understanding Priors in Bayesian Neural Networks at the Unit Level**. ICML, 2019.
- Mariia Vladimirova, Stéphane Girard, Hien D Nguyen, and Julyan Arbel. **Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions**. Stat, 2020.
- Mariia Vladimirova, Julyan Arbel, and Stéphane Girard. **Bayesian neural network unit priors and generalized Weibull-tail property**. ACML, 2021.
- Mariia Vladimirova, Julyan Arbel, and Stéphane Girard. **Dependence between Bayesian neural network units**. BDL workshop, NeurIPS, 2021.



- Theodore Papamarkou et al. **Position Paper: Bayesian Deep Learning in the Age of Large-Scale AI.** ICML, 2024.



- Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. **A primer on Bayesian neural networks: review and debates**. Statistical Science, 2024. https://arxiv.org/abs/2309.16314

Credits

- Blog post by Vincent Fortuin
  https://neptune.ai/blog/bayesian-deep-learning-needed-in-the-age-of-large-scale-ai
- Blog post by Piotr Januszewski
  https://neptune.ai/blog/bayesian-neural-networks-with-jax
- Slides on Laplace approximation by Kouroche Bouchiat
  https://bouchi.at/

References


Arbel, J., Dang, H.-P., Elvira, C., Herzet, C., Naulet, Z., and Vladimirova, M. (2023).
Bayes in action in deep learning and dictionary learning.
ESAIM: Proceedings and Surveys, 74:90–107.


Arbel, J., Pitas, K., Vladimirova, M., and Fortuin, V. (2024).
A primer on Bayesian neural networks: review and debates.
Statistical Science.


Bachoc, F. and Lagnoux, A. (2025).
Posterior contraction rates for constrained deep Gaussian processes in density estimation and classification.
Communications in Statistics - Theory and Methods, 54(3):774–811.


Bhattacharya, S., Liu, Z., and Maiti, T. (2020).
Variational Bayes neural network: Posterior consistency, classification accuracy and computational challenges.
arXiv preprint arXiv:2011.09592.


Bhattacharya, S. and Maiti, T. (2021).
Statistical foundation of Variational Bayes neural networks.
Neural Networks, 137:151–173.


Bos, T. and Schmidt-Hieber, J. (2022).
Convergence rates of deep ReLU networks for multiclass classification.
Electronic Journal of Statistics, 16(1):2724 – 2773.


Bouchiat, K., Immer, A., Yèche, H., Rätsch, G., and Fortuin, V. (2023).
Laplace-approximated neural additive models: Improving interpretability with Bayesian inference.
arXiv preprint arXiv:2305.16905.


Braun, A., Kohler, M., Langer, S., and Walk, H. (2024).
Convergence rates for shallow neural networks learned by gradient descent.
Bernoulli, 30(1):475–502.


Castillo, I. and Egels, P. (2024).
Posterior and variational inference for deep neural networks with heavy-tailed weights.


Castillo, I. and Randrianarisoa, T. (2024).
Deep Horseshoe Gaussian Processes.
arXiv preprint arXiv:2403.01737.


Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. (2021).
Laplace redux - effortless Bayesian deep learning.
In *Advances in Neural Information Processing Systems*.


Finocchio, G. and Schmidt-Hieber, J. (2023).
Posterior Contraction for Deep Gaussian Process Priors.
Journal of Machine Learning Research, 24(66):1–49.


Gal, Y. and Ghahramani, Z. (2016).
Dropout as a Bayesian approximation: Representing model uncertainty in deep learning.
In *International Conference on Machine Learning*.


Jantre, S., Bhattacharya, S., and Maiti, T. (2024).
Spike-and-Slab Shrinkage Priors for Structurally Sparse Bayesian Neural Networks.
IEEE Transactions on Neural Networks and Learning Systems.


Kohler, M. and Langer, S. (2021).
On the rate of convergence of fully connected deep neural network regression estimates.
The Annals of Statistics, 49(4):2231–2249.


Lee, K. and Lee, J. (2022).
Asymptotic Properties for Bayesian Neural Network in Besov Space.
In *Advances in Neural Information Processing Systems*.

MacKay, D. (1992).
A practical Bayesian framework for backpropagation networks.
Neural Computation, 4(3):448–472.

Murphy, K. P. (2023).
Probabilistic Machine Learning: Advanced Topics.
MIT Press.

Neal, R. (1995).
Bayesian learning for neural networks.
Springer.

Papamarkou, T., Skoularidou, M., Palla, K., Aitchison, L., Arbel, J., Dunson, D., Filippone, M., Fortuin, V., Hennig, P., Hernández-Lobato, J. M., Hubin, A., Immer, A., Karaletsos, T., Khan, M. E., Kristiadi, A., Li, Y., Mandt, S., Nemeth, C., Osborne, M. A., Rudner, T. G. J., Rügamer, D., Teh, Y. W., Welling, M., Wilson, A. G., and Zhang, R. (2024).
Position: Bayesian deep learning is needed in the age of large-scale AI.
International Conference on Machine Learning.

Pitas, K. and Arbel, J. (2023).
The fine print on tempered posteriors.
Asian Conference on Machine Learning.

Sharma, M., Farquhar, S., Nalisnick, E., and Rainforth, T. (2023).
Do bayesian neural networks need to be fully stochastic?
In *International Conference on Artificial Intelligence and Statistics*, pages 7694–7722. PMLR.

Vladimirova, M., Arbel, J., and Girard, S. (2021).
Bayesian neural network unit priors and generalized Weibull-tail property.
Asian Conference on Machine Learning.

Vladimirova, M., Girard, S., Nguyen, H. D., and Arbel, J. (2020).
Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions.
Stat.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019).
Understanding Priors in Bayesian Neural Networks at the Unit Level.
International Conference on Machine Learning.

Wolinski, P. and Arbel, J. (2025).
Gaussian Pre-Activations in Neural Networks: Myth or Reality?
Transactions on Machine Learning Research.